YEARBOOK OF PHYSICAL ANTHROPOLOGY 41:65–92 (1998)

Stature Estimation and Calibration: Bayesian and Maximum Likelihood Perspectives in Physical Anthropology

LYLE W. KONIGSBERG,[1] SAMANTHA M. HENS,[2]
LEE MEADOWS JANTZ,[1] AND WILLIAM L. JUNGERS[3]
[1]*Department of Anthropology, University of Tennessee Knoxville, Tennessee 37996–0720*
[2]*Department of Cell Biology and Anatomy, Johns Hopkins University, Baltimore, Maryland 21205*
[3]*Department of Anatomical Sciences, SUNY–Stony Brook, Stony Brook, New York 11794–8081*

KEY WORDS stature estimation; calibration; regression analysis

ABSTRACT Many applied problems in physical anthropology involve estimation of an unobservable quantity (such as age at death or stature) from quantities that are observable. Two of the more disparate subdisciplines of our discipline, paleoanthropology and forensic anthropology, routinely make use of various estimation methods on a case-by-case basis. We discuss the rationales for making estimations on isolated cases, taking stature estimation from femoral and humerus lengths as an example. We show that the entirety of our discussion can be placed within the context of calibration problems, where a large calibration sample is used to estimate an unobservable quantity for a single skeleton. Taking a calibration approach to the problem highlights the essentially Bayesian versus maximum likelihood nature of the question of stature estimation. On the basis of both theoretical arguments and practical examples, we show that inverse calibration (regression of stature on bone length) is generally preferred when the stature distribution for a reference sample forms a reasonable prior, while classical calibration (regression of bone length on stature followed by solving for stature) is preferred when there is reason to suspect that the estimated stature will be an extrapolation beyond the useful limits of the reference sample statures. The choice between these two approaches amounts to the decision to use either a Bayesian or a maximum likelihood method. Yrbk Phys Anthropol 41:65–92, 1998.

TABLE OF CONTENTS

Regression analysis is one of the most heavily used statistical methods in physical anthropology. Largely as a consequence of its ubiquity and wide availability in statistical packages and spreadsheets, regression analysis is typically applied by rote. In some cases this leads to analyses that, at least on the surface, appear strange. For example, we are used to seeing osteon counts as independent variables that predict the dependent variable age. Within an explanatory framework, it is difficult to see how osteons or any other "age marker" cause, even in part, age. We cannot assume that gaining osteons somehow causes one to get older. On the other hand, there is nothing that requires that we attach a causal model to regression analyses. If the goal in estimation is to obtain a minimum variance unbiased estimator and if we are willing to assume that our future cases come from the same distribution as the data on which we formed the regressions, then we can justify the regression of what will be unobservable (in future cases) on what will be observable. The logic behind this justification, which can be found in the calibration literature (Brown, 1982, 1993; Martens and Næs, 1991; Osborne, 1991), forms much of the basis for this paper.

In the general estimation problem, there are two broadly different approaches that we could apply. The first approach is to construct a maximum likelihood estimator, where we seek the posterior density for an estimate by maximizing the likelihood of observing a fixed data point (such as femur length) conditional on the unobservable value we are estimating (such as stature). The second approach is Bayesian, where we use the likelihood of observing the fixed data point but also allow for an informative prior for the estimate (again, in this case, for stature). Curiously, the second approach is very common in physical anthropology despite most physical anthropologists having very limited experience with Bayesian methods. One goal in writing this article is to show that, regardless of whether the inferential basis is clearly stated, many physical anthropologists have already applied Bayesian analyses and consequently need to become better acquainted with this branch of statistical theory. As another example of the ubiquity of Bayesian methods in physical anthropology, discussions of discriminant analyses have typically focused on concepts such as the prior and posterior probabilities of group membership. These are explicitly Bayesian concepts. Litton and Buck (1995) have recently made a strong argument for the adoption of Bayesian approaches in archaeology, while Lucy et al. (1996) have suggested a Bayesian approach be used in estimating age at death. We know of no general argument for the use of Bayesian approaches in physical anthropology.

We deal here exclusively with the problem of estimating stature from long-bone lengths, though most of the arguments we make can be applied to any estimation problem in physical anthropology. In this article, we contrast Bayesian and maximum likelihood approaches to stature estimation. We show that for some situations, notably most forensic cases, we can make a fairly strong Bayesian assumption that the case for which we are estimating an unobservable quantity comes from the same distribution as the reference sample. In contrast, in paleoanthropology and bioarchaeology we generally have no direct way of knowing that a case is sampled from the same distribution as the reference material. Because the assumptions for these two settings differ, we will find that the preferred method for estimation in each context also differs.

MEASUREMENT MODELS IN STATURE ESTIMATION

Before discussing the various methods for estimation of stature, it will be necessary for us to establish some minimal notation and terminology. We will refer to a dependent

variable, such as femur length, as the *y* variable and the independent variable stature as the *x* variable. This usage is at odds with the forensic anthropology literature, where stature is usually considered as a dependent variable whose value is predicted by an independent variable such as femur length. In contrast, in the allometry literature long-bone length is commonly considered as a dependent variable. In that literature, stature (or weight) is a measure of body size which may in part explain the size of a particular organ. Thus, it is natural that we should write stature as the *x* variable and long-bone length as the *y* variable, as the latter is a measure of organ size (individual bones being organs within the skeletal system). One can raise the literal argument that while long-bone length is an imprecise measure of the size of the femoral organ, so too is stature an imprecise measure of body size. From this viewpoint, both variables are partially dependent on some external overall (and unmeasured) variable of body size. We do not find this to be a very productive viewpoint, as it confuses the description and analysis of scaling with the problem of estimating an unobserved variable (in this case, stature). That there is a distinction between these two areas has been made clear in Smith's (1994) recent commentary (see also Seim and Sæther, 1983; Ricker, 1973).

We can further justify our use of long-bone length as a dependent variable and stature as an independent variable if we appeal directly to the calibration literature. To quote from Schéffe (1973:2):

> There are two related quantities $\mathscr{U}$ and $\mathscr{V}$ such that $\mathscr{U}$ is relatively easy to measure and $\mathscr{V}$ relatively difficult, requiring more effort or expense; furthermore the error in measurement of $\mathscr{V}$ is negligible compared with that of $\mathscr{U}$.

Following this description, Schéffe gives as one of his assumptions that $\mathscr{U}$ can be modeled as a linear function of $\mathscr{V}$. In our context, the "expensive" variable is stature, because in many cases stature cannot be obtained by any direct method. In contrast, long-bone length is "relatively easy to measure," since it can be obtained from any individual who is in a reasonable state of preservation.

Schéffe's comments regarding measurement error are also of import here. To address the problem of measurement error, we need to distinguish between two types of calibration: absolute vs. comparative. The distinction between these two forms of calibration is discussed in Osborne (1991) and the references therein. One of the clearest and most concise contrasts between absolute and comparative calibration is given in Rosenblat and Spiegelman (1981). They (Rosenblat and Spiegelman, 1981:329) describe absolute calibration as follows:

> A quick or nonstandard measurement technique is calibrated against a standard or defined measurement. . . In absolute calibration, the standard measurement is either known or made with negligible error.

On the other hand, in comparative calibration,

> One instrument or measurement technique is calibrated against another, with neither one being inherently a standard.

Under these definitions, stature estimation from long-bone length is clearly an example of absolute calibration. Although there is some error in the measurement or reporting of stature (Giles and Hutchinson, 1991; Willey and Falsetti, 1991), this error is negligible when compared to the uncertainty of stature measurement as provided from regression analyses (i.e., calibrations) of long-bone lengths.

METHODS OF STATURE ESTIMATION

Univariate-univariate methods

In this section we describe a number of different methods that can be used to estimate stature from the measurement of a single bone. Because we are estimating a single variable (stature) from a single measurement (e.g., long-bone length), we refer to this as a univariate-univariate problem. Later in the text we will consider multivariate-univariate methods, where multiple bones are used to estimate stature. We refer to this more general setting either as a multivariate-univariate problem or more simply as a multivariate problem. Here we consider five statistical methods for estimating stature: 1) regression of stature on a long-bone length, 2) regression of long-bone

length on stature followed by solving for stature, 3) major axis regression of stature on long-bone length, 4) reduced major axis regression of stature on long-bone length, and 5) use of a long-bone/stature ratio. With the exception of the second method, the other methods have been extensively discussed in the literature. Aiello (1992) and Sjøvold (1990) have given excellent reviews of ordinary least-squares regression of stature on long-bone length, major axis regression (MA), and reduced major axis (RMA) regression, while Rayner (1985) and Martin and Barbour (1989) have reviewed bivariate models in biomechanical and allometric applications, respectively. In addition, Feldesman et al. (1990) have suggested and carefully documented the use of a femur/stature ratio to estimate stature. Despite the antiquity of this method (see comments in Feldesman et al., 1990), the femur/stature ratio has received only rather limited discussion in the paleoanthropological literature.

The first stature estimator (regression of stature on long-bone length) could simply be called *regression*. Indeed, in a recent article Aykroyd et al. (1997) primarily refer to a comparable problem (regression of age on an indicator) as *regression*. They also refer to this estimator as *inverse calibration*, which is the term we use here. We prefer this latter appellation because it draws attention to the fact that we are regressing x on y rather than the usual y on x. The term *inverse calibration* was first used by Krutchkoff (1967). Unfortunately, there is not much standardization in the literature, as *inverse calibration* is also called "uncontrolled calibration" (Stuart and Ord, 1991), "direct calibration" (Afifi and Clark, 1996), and "natural or random calibration" (Brown, 1993). In any event, in what we refer to as *inverse calibration*, stature (x) is regressed on a long-bone length (y) so that the stature for an individual can be estimated by ordinary least squares (OLS) as

$$\hat{x}_i = \bar{x} + \beta_{xy}(y_i - \bar{y}). \quad (1)$$

In this equation, β_{xy} is equal to cov (x, y)/V_y, with cov(x, y) being the covariance between stature and femur length and V_y being the variance of femur length.

The second estimator (regression of long-bone length on stature followed by solving for stature) we will refer to as *classical calibration*. Krutchkoff (1967) originally used this term in reference to the longer history for this method, which dates back to at least the late 1930s (Eisenhart, 1939). Unfortunately, again there is little terminological standardization in the literature. Neter et al. (1985) refer to *classical calibration* as "inverse prediction," Brownlee (1965) refers to this method as "the use of the regression line in reverse," Stuart and Ord (1991) call it "conditional calibration," Brown (1993) calls it "controlled calibration," and Afifi and Clark (1996) call it "indirect calibration." Aykroyd et al. (1997) refer to "classical calibration" in their text, but some of their figures simply refer to this method as "calibration." Classical calibration (OLS regression of y on x, followed by solving for x) leads to the estimator

$$\hat{x}_i = \bar{x} + \beta_{yx}^{-1}(y_i - \bar{y}), \quad (2)$$

where β_{yx}^{-1} is equal to V_x/cov(x, y). Classical calibration has not seen much use within physical anthropology. Aykroyd et al. (1997) and Konigsberg et al. (1997) consider the method within the context of age estimation, while Rogers (1996) gives a specific example from histological aging. In a novel application, C. Stojanowski (unpublished) has used classical calibration to estimate site dates from cranial morphology. Hens et al. (1998) have specifically used classical calibration within the context of stature estimation.

The femur/stature ratio is an alternative estimator for stature which has been promulgated by Feldesman et al. (1990). They divide an observed femur length by the ratio of mean femur length to mean stature in order to obtain stature. We can equally well treat the ratio of average stature to femur length as a regression parameter to be multiplied by an observed femur length. The femur/stature estimator for stature is then

$$\hat{x}_i = \frac{\bar{x}}{\bar{y}} y_i. \quad (3)$$

This equation implies a regression line constrained so that it must pass through the origin and through the bivariate mean. Un-

less the femur exhibits isometry with stature, this will not be a least-squares line. Bissell (1992) refers to such a model as a "constant ratio" model.

The fourth and fifth methods (RMA and MA regression of stature on long-bone length) have been very extensively reviewed by Aiello (1992) and Sjøvold (1990). The reduced major axis regression is simply

$$\hat{x}_i = \bar{x} + \sqrt{\frac{V_x}{V_y}}\,(y_i - \bar{y}), \tag{4}$$

while the major axis regression is

$$\hat{x}_i = \bar{x} + \frac{\mathrm{cov}\,(x, y)}{\lambda_1 - V_X}\,(y_i - \bar{y}), \tag{5}$$

where λ_1 is the first eigenvalue of the variance-covariance matrix of stature and femur length (e.g., for calculation of the first eigenvalue see Sokal and Rohlf, 1981). There is an equivalent way of writing the RMA and MA regressions which is slightly more complicated but more general in that it extends easily to the multivariate (as opposed to bivariate) case. In this setting, we write λ (not subscripted) as the ratio of error variances for x and y. We then find the second (minor) eigenvector of either the variance-covariance matrix (for MA) or correlation matrix (for RMA) between x and y. We write these eigenvector elements as e_{x2} and e_{y2}. The RMA and MA equations are then

$$\hat{x}_i = \bar{x} - \sqrt{\lambda}\,\frac{e_{y2}}{e_{x2}}\,(y_i - \bar{y}) \tag{6}$$

where λ is equal to 1.0 for MA (i.e., the error variances for x and y are assumed to be equal), or the ratio of V_x to V_y for RMA (i.e., the ratio of error variances is assumed equal to the ratio of marginal variances). In the RMA case, the two elements of the second eigenvector are equal but with opposite signs, which shows the equivalence of equation 6 with equation 4. RMA and MA are both special cases of functional relationships, where both variables are presumed to contain measurement error. Sprent (1969) gives an extensive discussion and historical background to the general functional relationship problem, while Aiello (1992) can be consulted for a detailed treatment of MA and RMA in allometric studies. Bolfarine and Lima (1997) consider functional relationships within the context of calibration problems.

Materials

Throughout this paper we will use maximum femur length, maximum humerus length, and stature data from various sources in order to make our arguments. The specific samples we use are from the Terry collection (n = 852), World War II data (n = 1,042) (Trotter and Gleser, 1952), and the Forensic Databank (n = 159) (Moore-Jansen et al., 1994). The Terry and WWII data were taken from 80 column computer punch cards kept at the Bernard Becker Medical Library Archives (Washington University, St. Louis, MO) (for further description see Jantz, 1996). We also use data on 19 African Pygmies (Jungers, 1988). These individuals have statures which were estimated using the Fully (1956) technique. One individual of the 19 is missing a humerus, so we have substituted the missing measurement using an EM algorithm (Little and Rubin, 1987) that regresses humerus length on femur length and stature. Finally, we include A.L. 288–1 in some examples, taking her stature to be 1,050 mm (Geissmann, 1986) and her femoral length to be 281 mm (Jungers, 1982). The stature of 1,050 mm for A.L. 288–1 is an estimate rather than a known stature, but it was obtained "anatomically" rather than by any of the estimators discussed here (see Feldesman and Lundy, 1988).

Table 1 contains the summary statistics for these samples. There are some obvious differences between samples that are explained by sample composition. For example, the WWII sample is significantly taller than all other samples because it is composed entirely of males. Although for ease of use we present all regression equations here (Table 2 contains the parameters for regression of stature on femur length), we should point out that the necessary parameters for all five estimators (inverse calibration, classical calibration, femur/stature ratio, RMA, and MA) as well as their standard errors can be obtained from Table 1. While this may seem to be a trivial point, we raise it because it emphasizes that the sufficient summary multivariate statistics

TABLE 1. Means and variance-covariance matrices for stature, humerus, and femur for various samples[1]

	Stature (mm)	Humerus length (mm)	Femur length (mm)
WWII (n = 1,042)			
Stature (mm)	8,582		
Humerus length (mm)	1,668	444	
Femur length (mm)	2,422	544	861
	1,740	338	475
Terry (n = 852)			
Stature (mm)	4,603		
Humerus length (mm)	704	208	
Femur length (mm)	1,261	256	489
	1,708	325	456
Forensic (n = 159)			
Stature (mm)	10,077		
Humerus length (mm)	1,848	514	
Femur length (mm)	2,564	599	915
	1,725	329	465
Pygmy (n = 19)			
Stature (mm)	6,202		
Humerus length (mm)	1,078	218	
Femur length (mm)	1,503	285	453
	1,405	276	376
Total non-Pygmy (n = 2,053)			
Stature (mm)	7,270		
Humerus length (mm)	1,375	390	
Femur length (mm)	2,092	487	798
	1,725	332	466

[1] The variance-covariance matrices are shown in lower triangular form, while the rows below the matrices contain the means.

should be presented for multivariate studies (Konigsberg, 1991). Figure 1 shows the bivariate plots for the 2,072 humans. Because the data set is so large and would produce a bleary mess if we plotted all points, we have shown the 2,053 non-Pygmies by plotting only those points that fall outside of the fifth convex hull (as small dots) and then plotting every fifth convex hull in the "peel." A convex hull "is simply a closed convex polygon whose vertices are those points around the periphery of the configuration" (Green, 1981:3). Green (1981) gives a useful description of "peeling bivariate data" using convex hulls, and Holcomb and Konigsberg (1995) give an example of using convex hulls in 2-D morphometric analysis. To finish Figure 1, we plot the Pygmies as large open circles. Aside from some allometric differences between the Pygmies and non-Pygmies evident in the graph, the general impression is that these data do come from a multivariate normal distribution and that the relationships among measurements are linear on the linear scale. We have not done any formal statistical tests for linearity or normality. With this many cases, virtually any null hypothesis could be rejected regardless of whether the results had any substantive meaning. We also have not examined regression residuals, but, if the data are indeed multivariate-normal, then the residuals should also be normally distributed. The effect of predictable allometric differences between human groups on size estimation has been extensively covered in Ruff (1994) and Holliday and Ruff (1997). Hens et al. (1998) consider the more dismal prospect of interspecific allometric differences that may not be predictable.

A worked example of stature estimation

It would be useful at this point to present a brief worked example of stature estimation from a single bone. For this example we use the parameters from Table 2 (based on 2,053 humans) to estimate stature for A.L. 288–1. When "Lucy's" femoral measurement of 281 mm is entered into the regression equations from Table 2, her estimated stature from inverse calibration is (to the nearest millimeter) 1,239 mm, while from classical calibration her estimated stature is 1,081 mm. RMA and MA give estimates of 1,166 and 1,097, respectively, while the ratio estimate is 1,097. The estimates from classical calibration, MA, and the ratio estimator are all fairly close to the anatomical reconstruction of 1,050 mm, while the inverse and RMA estimates are clearly too tall. In addition to these point estimates from each method, it would be useful to have a confidence interval around each estimate. Table 2 lists the square roots of the integrated mean squared errors (IMSE). The asymptotic 95% confidence interval around each estimate is plus and minus 1.96 times the IMSE. This yields intervals for inverse calibration, classical calibration, RMA, MA, and the ratio estimator of from 1,156–1,323, from 986–1,176, from 1,080–1,252, from 1,004–1,190, and from 937–1,142 mm, respectively. It is interesting to note that inverse calibration and RMA do not include the anatomical estimate of 1,050 mm within their 95% confidence intervals. Later in the text we discuss the meaning of IMSE and the way it is calculated. It may appear

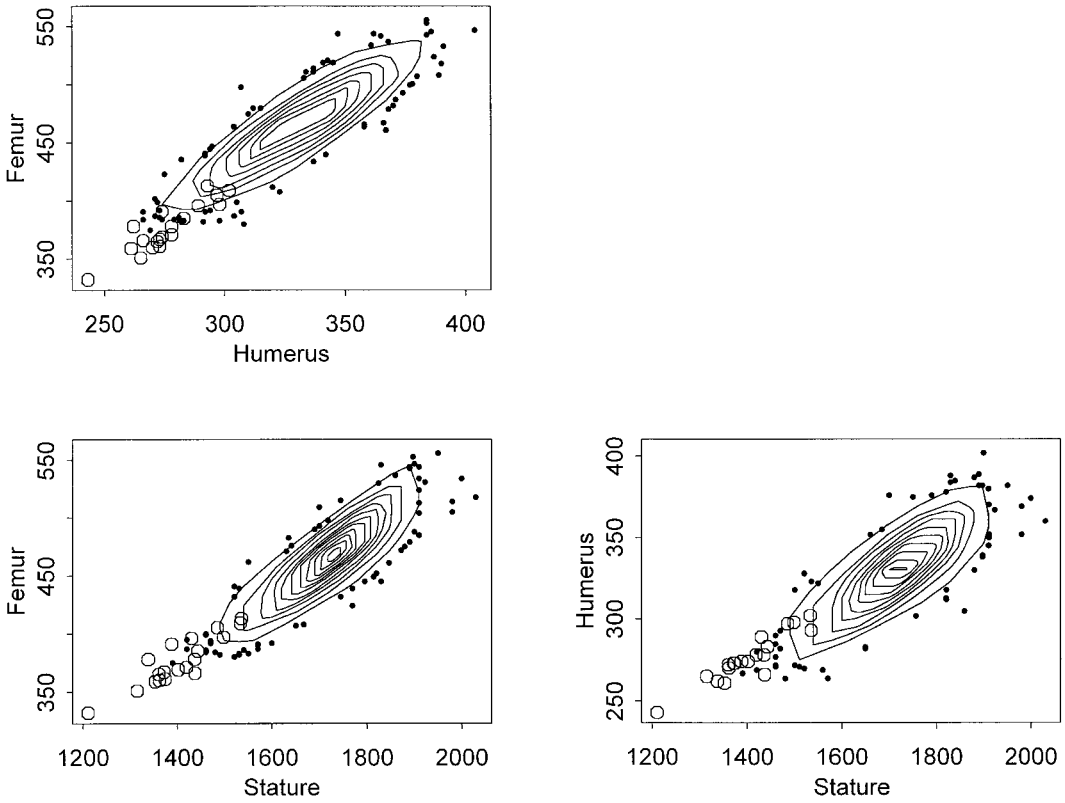

Fig. 1. Bivariate plots of femur length against humerus length, femur length against stature, and humerus length against stature for 2,072 humans. The large open circles represent 19 Pygmies, while the small points represent non-Pygmies who fall outside the fifth convex hull. The polygons are the fifth, tenth, fifteenth, etc., convex hulls for the non-Pygmies.

TABLE 2. Regression equations for predicting stature (in millimeters) from maximum femoral length (in millimeters)[1]

Method	Intercept	Femur	$\sqrt{\text{IMSE}}$*
Inverse calibration	502.95	2.6209	42.45
Classical calibration	104.60	3.4751	48.67
RMA	317.80	3.0179	43.73
Ratio	0.0	3.6993	52.11
MA	144.91	3.3886	47.51

[1] Based on 2,053 humans (Terry, WWII, and Forensic Databank).
*Square root of the integrated mean squared error, where integration is across the normal distribution with a mean of 1,725 and standard deviation of 85 mm (the distribution for stature in the sample of 2,053 humans).

curious at this point that our interval from inverse calibration does not take account of the large divergence of A.L. 288–1's femur from the modern human average of 466 mm (in contrast see Boldsen, 1984; Giles and Klepinger, 1988). However, because we have a very large reference sample, the additional variance term due to distance from the long-bone mean is negligible. Similarly, the curved bands that form confidence intervals in classical calibration (Carroll et al., 1988; Fisch and Strehlau, 1993) will become simple linear intervals if the reference sample is large enough.

BAYESIAN AND LIKELIHOOD JUSTIFICATIONS FOR THE ESTIMATORS

The classical calibration approach can be justified on the grounds that it is the maximum likelihood estimator (see Graybill, 1961). One way to appreciate this fact is to write the posterior probability for stature as proportional to the product of the likelihood with a locally uniform prior. In this case, the likelihood itself is proportional to the probability that an individual of a specified stature would have a long bone whose length

was identical to that which had been measured in a test case. For example, we might ask what the probability is that an individual who was 1,050 mm tall would have a femoral length that was 281 mm (as in A.L. 288–1). We would then use the regression information from a reference sample to find the probability distribution across possible statures. This is shown in Figure 2 (upper part), where we have used the combined Terry, WWII, and forensic data to obtain the likelihood. The likelihood is found by predicting femoral length from stature and then finding the probability of obtaining the predicted femoral length if the actual femoral length was 281 mm. This probability comes from a normal distribution with the mean equal to the predicted femoral length and the variance equal to the square of the standard error of estimate from the regression of femur length on stature. The estimated stature for A.L. 288–1 from classical calibration (shown as the posterior distribution in the upper right of Fig. 2) is 1,081 mm, with a standard error of 49 mm. Konigsberg and Frankenberg (1994) show how combining this likelihood with a uniform prior (as in the upper part of Fig. 2) leads algebraically to the classical estimator (regression of femur on stature followed by solving for stature), though their context is that of age rather than stature estimation.

In contrast to classical calibration, inverse calibration has generally been motivated by the argument that it produces a Bayes estimator (Hoadley, 1970; Hunter and Lamboy, 1981; Lwin and Maritz, 1982). In this context, the posterior probability for stature is proportional to the product of the likelihood with an informative prior distribution for stature. We show this in the lower part of Figure 2, where the prior is the normal distribution for stature from the combined Terry, WWII, and forensic data. This produces a posterior distribution for stature with a higher mean and which is more concentrated (i.e., has lower variance). The estimated stature is 1,239 mm with a standard error of 42 mm. Because the classical calibration approach takes an uninformative prior for stature, the prediction interval (with a 49 mm standard error) is (necessarily) larger than that from inverse calibration. Again, Konigsberg and Frankenberg (1994) show that combining the likelihood with an informative prior from the reference sample (as shown in the lower part of Fig. 2) leads algebraically to the inverse estimator.

The femur/stature ratio, RMA, and MA can all be justified as maximum likelihood estimators. In the case of the ratio estimator, the justification is the same as for classical calibration, with the restriction that the regression line of femur on stature must pass through the origin. RMA and MA are more complicated to justify on a likelihood basis. Our discussion follows Graybill (1961). Because there is truly no independent (x) or dependent (y) variable in the symmetric RMA and MA, we will use the symbols f and s for observed femur length and stature and F and S for the measurements that would be observed if there were no error. The ratio of measurement error variances for femur and stature we will write (without a subscript) as λ, reserving the subscripted symbol lambda for eigenvalues. The unobserved (i.e., measured without error) femur length for the ith individual can be estimated as

$$\hat{F}_i = \frac{f_i + \lambda\hat{\beta}(s_i - \hat{a})}{1 + \lambda\hat{\beta}^2}, \qquad (7)$$

where $\hat{\beta}$ and $\hat{\alpha}$ are the regression parameters from RMA or MA, and, as before, lambda is equal to 1.0 for MA and to the ratio of stature to femur variance for RMA. The kernel of the log-likelihood for the ith case is then

$$\ln LK = -\frac{1}{2}\left(\frac{(s_i - \alpha - \beta F_i)^2}{\sigma_F^2} + \frac{(f_i - F_i)^2}{\sigma_S^2}\right) \qquad (8)$$

where σ_F^2 and σ_S^2 are the error variances for the unobserved femur and stature lengths. Graybill (1961) gives an estimator for these variances in his Theorem 9.1, but a simple way to find them is to note that they are functions of the minor (second) eigenvalue of the correlation matrix for RMA or the variance-covariance matrix for MA. For RMA, the error variances are equal to the second eigenvalue times the appropriate marginal variance, while for MA the error variances are equal to the second eigenvalue. Rather than repeat Figure 2 with a different likelihood, we show in Figure 3 how

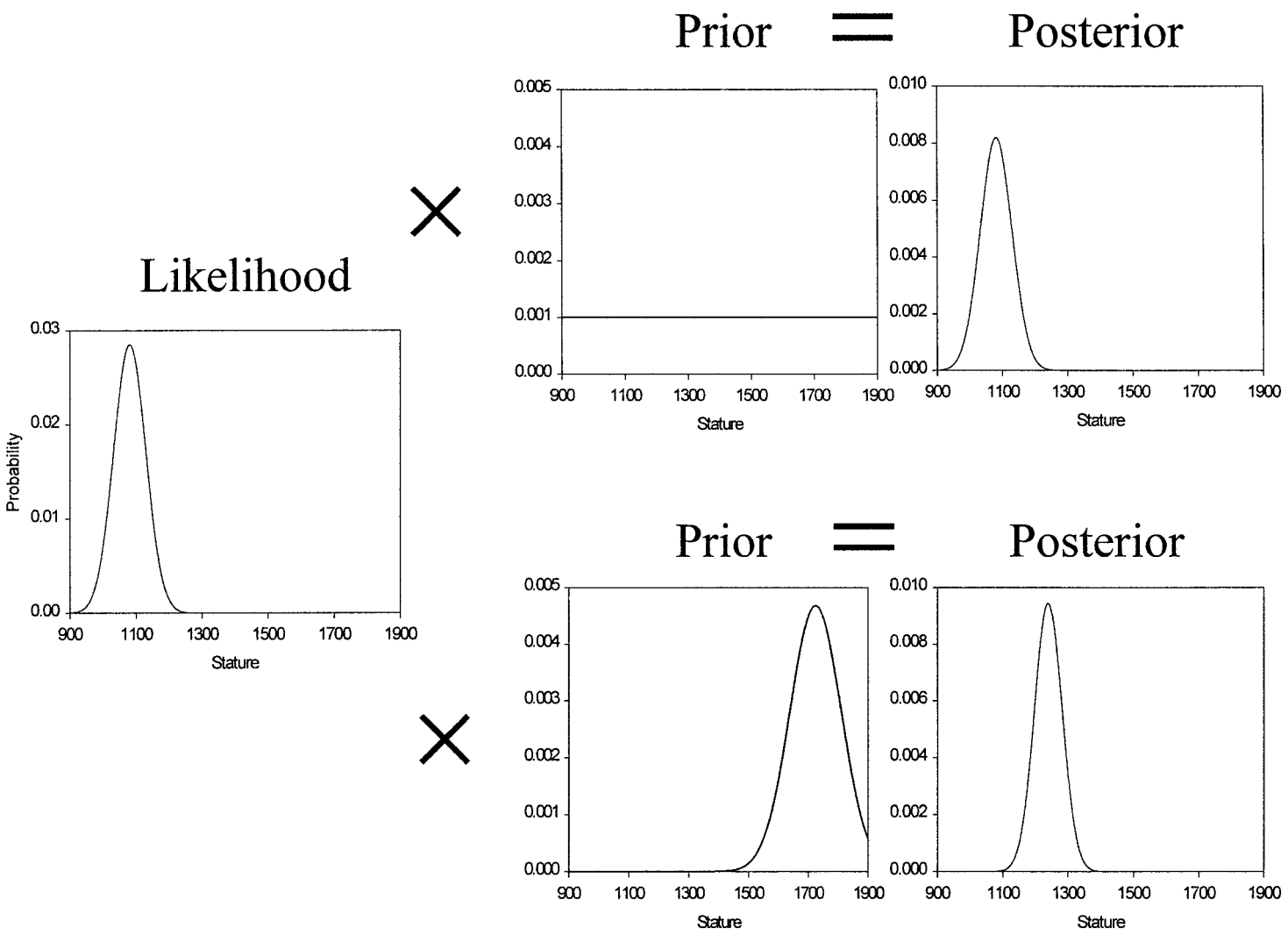

Fig. 2. Comparison of maximum likelihood (classical calibration) and Bayesian (inverse calibration) methods for estimating stature. The single graph to the left shows the likelihood for stature (in millimeters) conditional on a femur length of 281 mm. The likelihood was found using 2,053 individuals from the Terry, WWII, and Forensic Databank samples. The upper right two graphs show the maximum likelihood estimate, taking a uniform prior to find the posterior density for stature conditional on the femur length of 281 mm. The lower right two graphs show the Bayesian estimate, taking an informative prior for stature from the reference collection to find the posterior density.

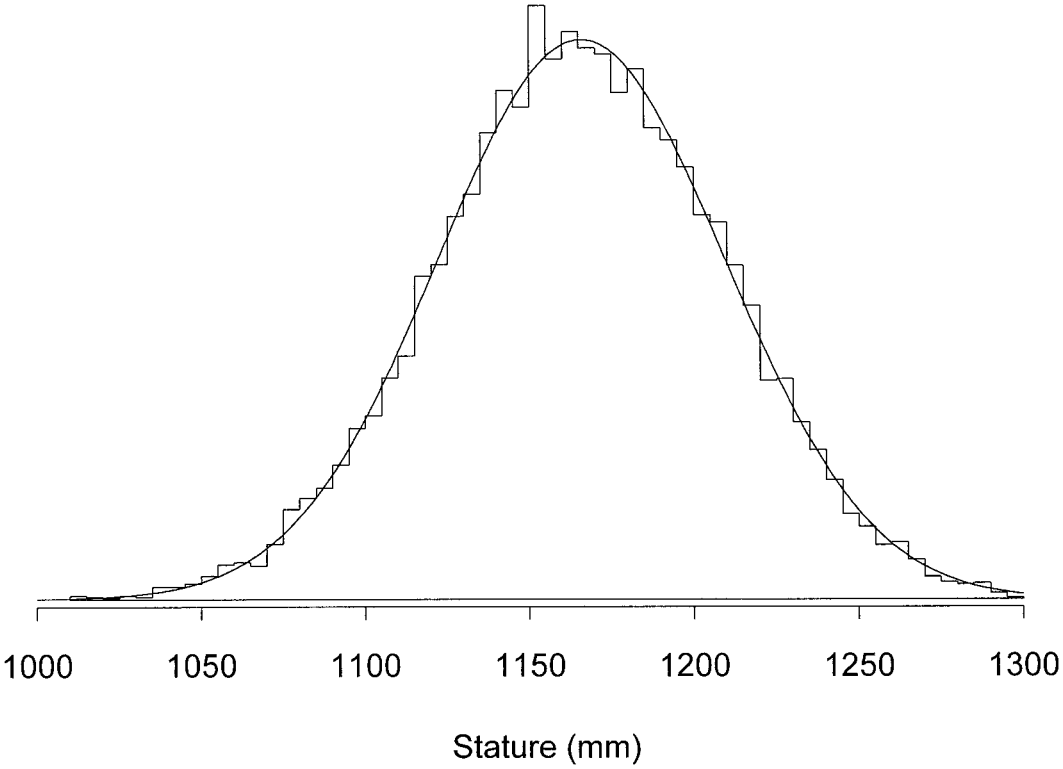

Fig. 3. Simulation of stature from the posterior density in RMA, conditioning on a femur length of 281 mm, and using individuals from the Terry, WWII, and Forensic Databank samples for the likelihood. The smooth line is the posterior density, while the histogram represents 10,000 draws via adaptive rejection sampling (Gilks and Wild, 1992). The 95% highest posterior density for stature from this graph is between 1,079.0 and 1,250.7 mm, with the average stature at 1,165.6 mm. The comparable figures derived using equations from Table 3 are a 95% confidence interval of from 1,080.1–1,251.6 mm around an estimate of 1,165.8 mm.

TABLE 3. Conditional variance, bias, and integrated mean squared error (IMSE) for five methods of stature estimation

Method	Conditional variance	Bias	IMSE
Inverse calibration	$V_x r^2 (1 - r^2)$	$(\overline{x} - x)(1 - r^2)$	$V_x(1 - r^2)$
Classical calibration	$V_x(r^{-2} - 1)$	0	$V_x(r^{-2} - 1)$
Reduced major axis	$V_x(1 - r^2)$	$(\overline{x} - x)(1 - r)$	$2V_x(1 - r)$
Femur/stature ratio	$\left(\frac{\overline{x}}{\overline{y}}\right)^2 V_y(1 - r^2)$	$(\overline{x} - x)\left(1 - \beta_{yx}\left(\frac{\overline{x}}{\overline{y}}\right)\right)$	$V_x - 2\frac{\overline{x}}{\overline{y}}\text{cov}(x, y) + \frac{\overline{x}^2}{\overline{y}^2} V_y$
Major axis	$\beta_{MA}^2 V_y (1 - r^2)$	$(\overline{x} - x)(1 - \beta_{yx}\beta_{MA})$	$V_x - 2\beta_{MA}\ \text{cov}(x, y) + \beta_{MA}^2 V_y$

equation 8 (and its first derivative with respect to stature) can be used to simulate statures from the posterior density conditional on an observed femur length. For this example, we use a femur length of 281 mm and the RMA model. The draws from the posterior density were made using adaptive rejection sampling (Gilks and Wild, 1992). The agreement between this density and summary statistics obtainable from Table 3 is, of course, quite high (see caption for Fig. 3). Although Jolicoeur (1968, 1990) has given confidence intervals for RMA and MA slopes, to our knowledge confidence intervals for femur/stature ratio predictions have not previously been discussed. The simulation that we have illustrated in Figure 3 could be used to construct intervals for any likelihood-based method, and it is relatively

easy to adapt the simulation to small-sample settings by including a draw on the regression parameters prior to simulating from the predictive distribution (see Gelman et al., 1995).

A fine but important point from this discussion is that even inverse calibration becomes a maximum likelihood estimator (MLE) if the implied measurement model is correct. Inverse calibration (regression of stature on femur length) is the MLE when only stature is measured with error. In this case, "inverse calibration" is really just traditional regression (as in Aykroyd et al., 1997). This is largely an academic point, as we will show below that statistical criteria can be used to select an estimator without having to pay attention to the sources of measurement error.

Mean squared errors, Pitman's measure of closeness, and stochastic dominance for the estimators

A number of different criteria could be used to decide which of the stature estimators is optimal. We primarily use the mean squared error (MSE) here. Feldesman and Fountain (1996) also use the mean absolute deviation (MAD) and Pitman's measure of closeness (PMC) as criteria for evaluating estimators. We do not use MAD here, as it does not materially alter our decisions regarding which of the estimators are desirable. We do use PMC both in its pairwise and simultaneous forms to compare estimators. We warn the reader, however, that there is considerable controversy surrounding the use of PMC (see in particular Robert et al. (1993) and the comments following that article). In particular, the PMC is subject to a number of paradoxes and anomalies, not the least of which is its intransitivity (see Chapter 3 of Keating et al., 1993). Because both PMC and MSE have been criticized on a number of fronts, we also consider stochastic dominance (Hwang, 1985) as a basis for selecting estimators.

The MSE is the average squared term of an estimator around its true value. For an unbiased estimator, the MSE reflects only uncertainty of the estimate. For a biased estimator, MSE reflects both uncertainty and a squared term due to bias. Consequently, the mean squared error must include both the conditional variance around the biased estimate and the square of the bias. These relationships are shown graphically in Figure 4 for two estimators, the classical and inverse calibrations. In this figure we assume that an individual was actually 1,050 mm tall, as has been suggested for A.L. 288–1. The regression of femur length on stature in our large reference sample of modern humans indicates that an individual 1,050 mm tall would have a femur length of 272 mm, with a standard error of 14 mm. Interestingly, this is close to A.L. 288–1's actual femur length of 281 mm. Figure 4A shows the 95% confidence interval for the expected femur length conditional on a stature of 1,050 mm. If classical calibration is now used to estimate stature,

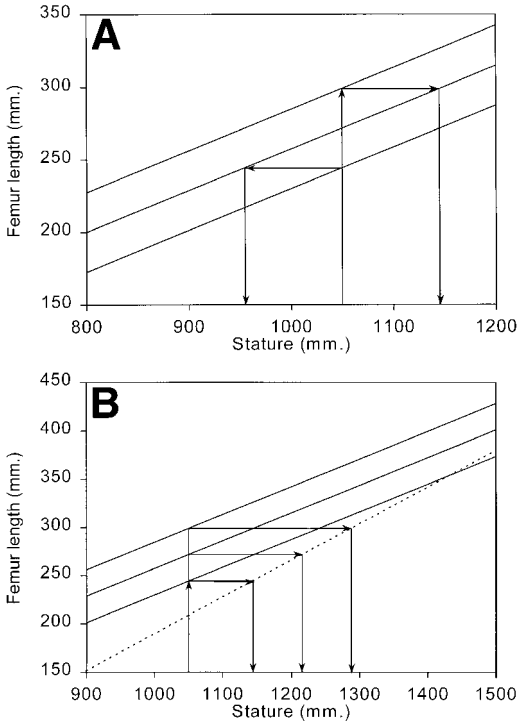

Fig. 4. Depiction of conditional variances for stature estimates from classical calibration (**A**) and inverse calibration (**B**). Each panel shows the variance of stature estimates for a case where the stature was actually 1,050 mm. In both panels, the three diagonal lines are the upper and lower 95% confidence limits around the estimated femur length given stature. These lines are taken from the data on 2,053 individuals referenced in Fig. 1. In **B,** the dashed line is from the regression of stature on femur length.

we need to project points back to this regression line and then project them down onto the x-axis. This produces a 95% confidence interval for stature of from (about) 955–1,145 mm with a width of 190 mm. The uncertainty represents the conditional variance. In Figure 4B, with inverse calibration, the 95% confidence interval is smaller (from 1,144–1,288 mm for a width of 144 mm), but this interval does not include the actual stature.

In Table 3 we present the large sample conditional variance for each estimator and the bias term. For classical and inverse calibration, these results have been presented numerous times (Berkson, 1969; Halperin, 1970; Martinelle, 1970; Shukla, 1972). In this table, the conditional variance indicates the variance of the estimator for a fixed (unobservable) stature, as in our previous graphical example. As a further example, if the actual stature were 1,600 mm, then (from the regression of femur on stature) this would produce a femur length from a normal distribution with a mean at 430 mm and standard deviation of 14 mm. The conditional variance is then the variance of the estimates (over many replications) that would be obtained using each estimator. By the parallel axis theorem, the MSE at a true value of stature is equal to this conditional variance plus the square of the bias of the estimator at the given (true) stature. The third column in Table 3 lists the integrated mean square error (IMSE). Ordinarily, this integration is across the empirical distribution for x (stature), so that IMSE is the same as the MSE that would be reported for a total sample (for a discussion of MSE and IMSE in the calibration context, see Brown, 1979).

A few salient facts are immediately apparent from Table 3. First, provided allometric scaling is preserved throughout, classical calibration produces an unbiased estimate through the entire range for stature. Except for the trivial case where the true stature is at the mean, inverse calibration and reduced major axis both give biased estimates of stature if the correlation between a long bone and stature is less than 1.0. The bias is greater for inverse calibration than for reduced major axis. Again with the exception of the trivial case at the mean, the femur/stature ratio will generally give biased estimates. The one case in which this is not true occurs if the long bone scales isometrically with stature, at which point the long-bone/stature ratio yields unbiased estimates for stature. Under isometry, β_{yx} equals $\bar{y}/\bar{x}$, so that the quotient of these terms (shown in the bias for femur/stature ratio in Table 3) will equal 1.0 and will cancel in the bias term. Finally, major axis regression always gives biased estimates (except at the mean stature), but the bias may be fairly trivial. For decreasing ratios of the variance of long-bone length to the variance of stature, bias will also decrease in the major axis regression. This is only an asymptotic result. At the limiting case where the ratio of the variances has gone to zero, the covariance between long-bone length and stature must also be zero, and the major axis slope is then undefined.

In order to obtain the MSE at a particular value of x (stature), we also need to consider the conditional variances. In the vicinity of the mean stature, where all methods are nearly unbiased, inverse calibration gives the lowest conditional variance, reduced major axis gives an intermediate value, and classical calibration gives the highest conditional variance of the three methods. It is on this basis that inverse calibration has often been favored, for, if there is no extrapolation, inverse calibration gives the lowest MSE. This is shown in Figure 5, where the expected mean squared errors are plotted against stature for each of the methods using the statistics from the sample of 2,053 individuals from the Terry, WWII, and forensic data. Figure 5A shows the MSE for inverse calibration, classical calibration, and reduced major axis regression of stature on femur. Figure 5B shows the MSE for classical calibration, femur/stature ratio, and major axis regression of stature on femur. In the first part of the graph (Fig. 5A), inverse calibration has a lower MSE than classical calibration whenever the true stature is between 1,595 and 1,855 mm (inclusive). RMA has a lower MSE than classical calibration whenever the true stature is between 1,542 and 1,908 mm. Inverse calibration has a lower MSE than does RMA when the true

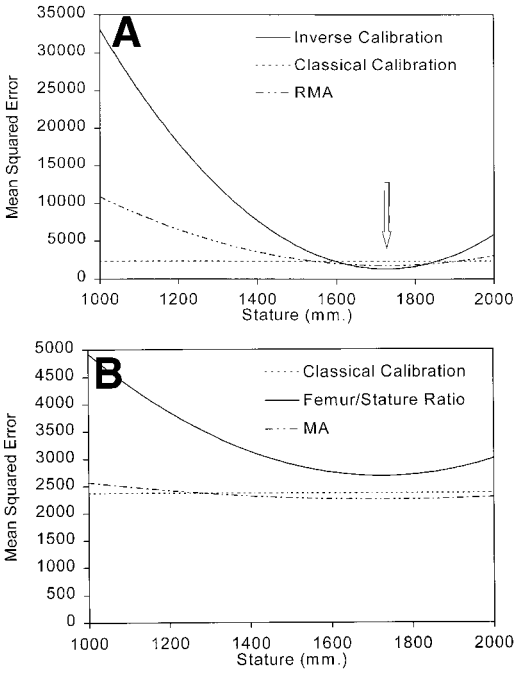

Fig. 5. Theoretical mean squared errors (MSE) for estimated stature conditional on true stature. **A:** The MSEs for inverse calibration, classical calibration, and reduced major axis regression (RMA). The vertical arrow represents the reference sample mean stature of 1,725 mm. **B:** The MSEs for classical calibration, the femur/stature ratio estimator, and major axis regression.

stature is between 1,625 and 1,825 mm. In the second graph (Fig. 5B), it is clear that there is little basis in the MSEs to pick between major axis regression and classical calibration. Both have fairly similar values for MSE, and their lines cross. In contrast, the femur/stature ratio always has a higher MSE than does classical calibration because the femur/stature ratio method is actually a special case of classical calibration (Hens et al., 1998). When a long bone scales isometrically with stature, then a long-bone/stature ratio method will coincide with the classical calibration method. In this sense, then, the femur/stature ratio can be as good as classical calibration if the femur actually scales isometrically with stature, but the ratio method can never be a better predictor of stature. From a decision-theoretic standpoint, the femur/stature ratio is MSE-inadmissible because it is dominated throughout by the classical calibration approach. In Discussion we will show that when the femur scales isometrically with stature, the femur/stature ratio approach is the same as classical calibration.

The MSE has been previously criticized as a basis for comparing classical and inverse calibration (Keating and Mason, 1991; Keating et al., 1993; Williams, 1969) because the unconditional MSE is infinite for classical calibration. This is true because, as the slope of the long bone regressed on stature approaches zero, the MSE from classical calibration gets very large, and at a slope of zero the MSE is infinite. It has frequently been pointed out that this problem can be avoided by using a truncated version of classical calibration (i.e., one in which the slope is not near zero). The truncated version of classical calibration has a finite MSE. The exclusion of a slope near or equal to zero has also been referred to as a conditional approach (e.g., see Lee, 1991), as it requires that the classical calibration proceed after conditioning on the slope not equaling zero. As the slope is certainly not near zero for the regression of femur length on stature, use of the truncated (or conditional) MSE should not cause concern. In any event, we have also considered PMC, as it has previously been used in comparisons of classical and inverse calibration (Halperin, 1970; Keating and Mason, 1991). PMC for a pairwise comparison of two estimators is the probability that one estimator will be closer to the true value than will the alternative estimator. *Closeness* refers to the Euclidean distance between the estimator value and the true value. For a simultaneous comparison of estimators, PMC is the probability that a given estimator is closer to the true value than is any other estimator under consideration. We use a Monte Carlo procedure to draw PMC curves, but we first compare this Monte Carlo PMC to the known analytical case of pairwise comparison between inverse vs. classical calibration (after Halperin, 1970). As the agreement between the Monte Carlo and the analytical solution is nearly perfect, we use the Monte Carlo for other applications where analytical solutions are not readily available. The Monte

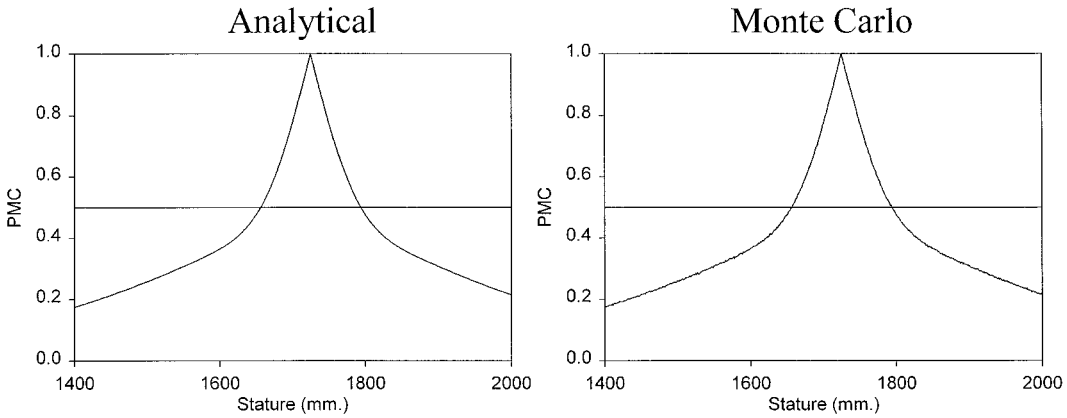

Fig. 6. Comparison of inverse calibration and classical calibration using Pitman's measure of closeness (PMC). The left graph shows this comparison using analytical results from Halperin (1970), while the right graph shows the comparison from a Monte Carlo procedure. PMC in these graphs is the probability that inverse calibration is (Euclidean) closer than classical calibration to the true value of stature. Portions of the graph that lie above the horizontal line at 0.5 consequently indicate regions where inverse calibration is PMC preferred.

Carlo procedure is quite simple. As we have a large calibration sample, the prediction interval for femur length conditional on stature is virtually a constant width. We simulate 100,000 replicates of femoral lengths at fixed statures in 1 mm increments between 1,400 and 2,000 mm. The replicates have their expectation taken from the regression of femur on stature and a variance equal to the squared standard error of estimate from this regression. We then use the femoral lengths to estimate stature and compare these estimates to the known stature or to other estimators. The Monte Carlo PMC is then the proportion of times out of 100,000 that the inverse calibration estimate is closer than the classical calibration estimate to the true value of stature.

Figure 6 shows the analytical solution and the Monte Carlo results for the PMC comparing inverse and classical calibration. These curves are virtually indistinguishable, and both indicate that, on the criterion of closeness, inverse calibration would be preferred over classical calibration if the actual stature was between 1,657 and 1,793 mm (inclusive). This is the region in which the probability that the inverse calibration estimator is closer to the actual stature than is the classical calibration estimator rises above 0.5. Figure 7 shows all of the pairwise PMC curves from the Monte Carlo procedure. This figure can be used to decide which estimator in a pairwise comparison would be preferable (on the basis of "closeness") both in the vicinity of the mean stature (as we presume might be the case for the forensic context) and away from the mean (as might be the case in paleoanthropology). In all pairwise comparisons between inverse calibration (regression of stature on femur length) and competing estimators, inverse calibration would be preferred in the vicinity of the mean. Conversely, though, away from the mean stature, inverse calibration is always the least desirable option (i.e., is "farthest" from the true stature). In the pairwise comparisons against classical calibration, classical calibration is always preferred when the actual stature is away from the mean (although this preference is slight in regard to major axis regression). Indeed, the comparison of classical calibration with the femur/stature ratio method shows that classical calibration always gives a "closer" estimate than the ratio does. Thus, in addition to the femur/stature ratio being MSE-inadmissible relative to classical calibration, it (femur/stature ratio) is also PMC-inadmissible. Figure 7 also shows that the femur/stature ratio is PMC-inadmissible when compared to major axis regression.

Figure 8 shows the "simultaneous" Monte Carlo PMCs for the "best" and "worst" cases.

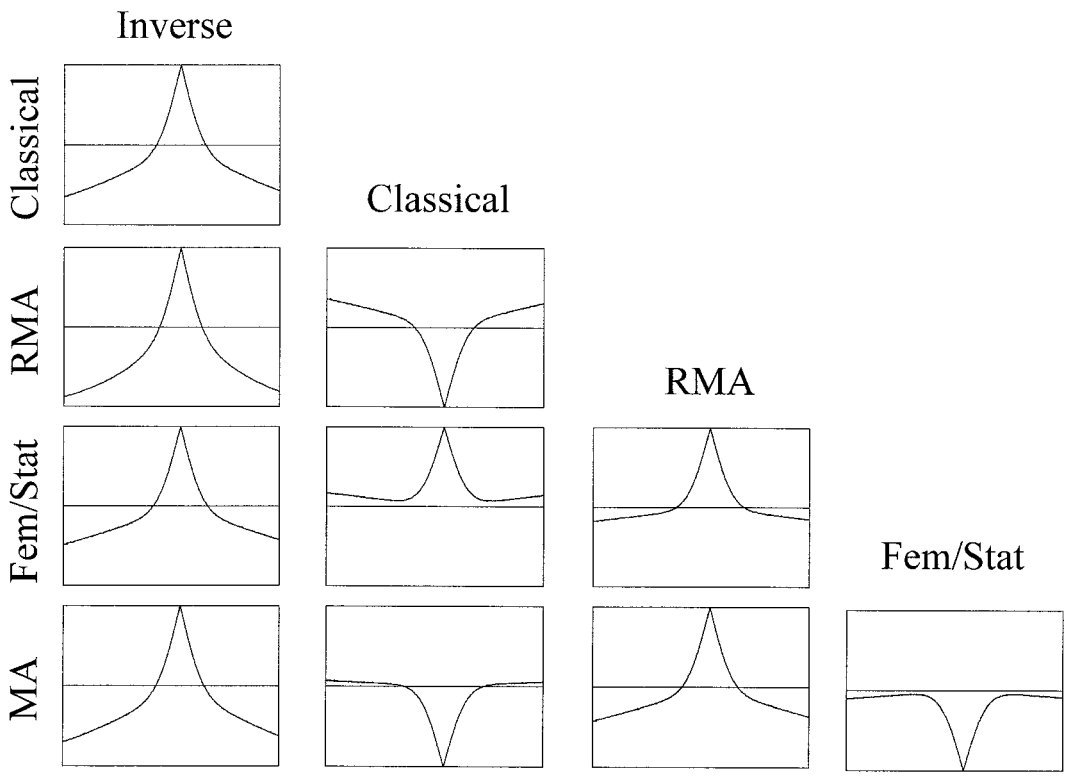

Fig. 7. Graphs of the pairwise Pitman measures of closeness (PMC) for the five stature estimators drawn using a Monte Carlo procedure. Axes are as in Fig. 6 but have been suppressed for legibility. Regions where the curves lie above the horizontal line at 0.5 represent areas where the column-labeled estimator is Pitman closer than the row-labeled estimator. For example, the graph in the third row and second column shows that classical calibration is always Pitman closer than the femur/stature ratio.

In the graph for the "best" estimator, the lines show the proportion of times out of 100,000 replicates at each stature that each estimator was Euclidean closest to the true stature value. Similarly, the graph for the worst estimator shows the proportion of times that each estimator was furthest from the true value of stature. Again, in the vicinity of the mean, inverse calibration is the preferred estimator relative to all other estimators. Away from the mean, the order of preference (i.e., relative closeness to the actual stature) places the femur/stature ratio as the best, reduced major axis as the second best, major axis as the third best, and classical calibration as the worst estimator. We disregard inverse calibration in this ranking, as its simultaneous PMC curve crosses the other estimators' curves. If we look instead at the simultaneous PMC worst estimator, only femur/stature ratio and inverse calibration appear in the graph. This occurs because the other three estimators (reduced major axis, major axis, and classical calibration) are often intermediate in terms of closeness, but they are never the worst (farthest from the actual value). The femur/stature ratio is always the worst estimator in the vicinity of the mean, while it vies with inverse calibration for this dubious honor away from the mean.

The results we have presented so far based on MSE and PMC would certainly be open to question by many statisticians. Use of MSE as a basis for selecting an estimator is often criticized because there is no natural justification for this particular loss function and because it may place too great a weight on values of x that are unlikely to occur (i.e., are far from the mean). PMC in turn is often

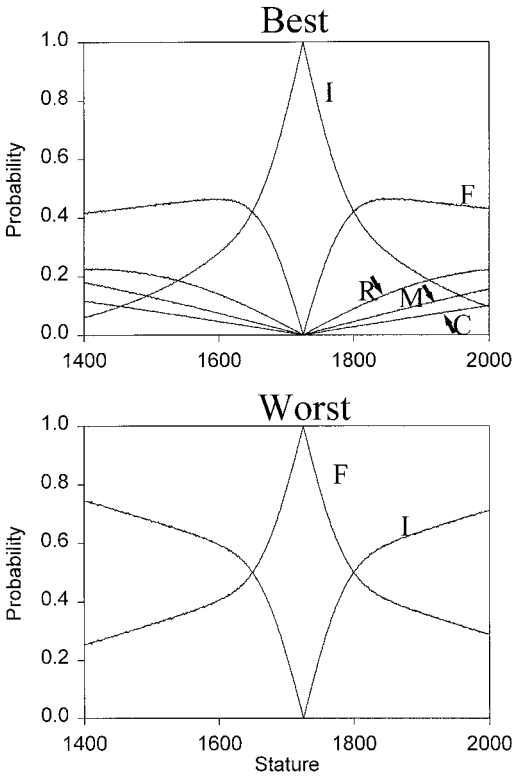

Fig. 8. Monte Carlo simultaneous Pitman measures of closeness for inverse calibration (I), classical calibration (C), reduced major axis (R), major axis (M), and femur/stature ratio (F). The top graph shows the probability that an estimator is the best (Euclidean closest), while the bottom graph shows the probability that an estimator is the worst (Euclidean furthest). Only femur/stature ratio and inverse calibration are shown in the "worst" graph, as the other three estimators have probabilities near zero of being "the worst."

criticized because of the paradoxes it frequently raises. In fact, Casella and Wells (1993) have written that the PMC "is an inherently flawed criterion that should not be the basis of evaluation of statistical procedures (1993:70)." We consequently also consider stochastic dominance (Hwang, 1985). The only major criticism of this method for evaluating estimators is that it may not lead to a clear choice among estimators unless one of them is phenomenally bad. Blyth (1993) noted that, in pairwise comparisons of estimators, "only for the most trivial, mismatched comparison would we be so lucky as to have stochastic domination decide between the two (1993:74)."

In using stochastic domination, we make comparisons between estimators by assessing the probability that an estimator is within some fixed Euclidean distance of the true value. An estimator stochastically dominates another one if the first estimator always has a higher probability of being within the fixed distance from the true value. To evaluate stochastic domination, we again use the Monte Carlo procedure previously described, where we find the proportion of times out of 100,000 replicates at each stature that an estimate is within some fixed distance of the actual stature. Figure 9 plots the probability across stature that each estimator will be within 72 mm of the true value. We selected this value so that inverse calibration would have a coverage of 95% at the mean stature. By *coverage* we mean that, if the true stature is at the mean, there is a 0.95 probability that inverse calibration would estimate stature within the 14.4 cm surrounding the mean. This concept is not identical with that of a confidence interval, because for this particular interval (of 14.4 cm) we have conditioned on the true stature being at the mean. Figure 9 shows that, like the MSE and PMC evaluations, inverse calibration would be the preferred estimator if stature is near the mean. Away from the mean, classical calibration, major axis regression and the femur/stature ratio would be preferred over reduced major axis and inverse calibration. Unlike the MSE and PMC comparisons, it is not valid to consider where lines cross in order to decide which estimators to use. By making the coverage very small, say 1 mm, we can substantially reduce the useful range of inverse calibration. Figure 9 shows the useful range for inverse calibration as quite wide only because we have taken a very broad coverage (14.4 cm). Stated another way, we cannot decide whether to use inverse calibration or some other estimator unless we make some statement about desired coverage. The comparisons between major axis, classical calibration, and the femur/stature ratio are interesting because they show cases of stochastic domination, where major axis regression is dominant to classical calibration, which in turn is dominant to the femur/stature ratio.

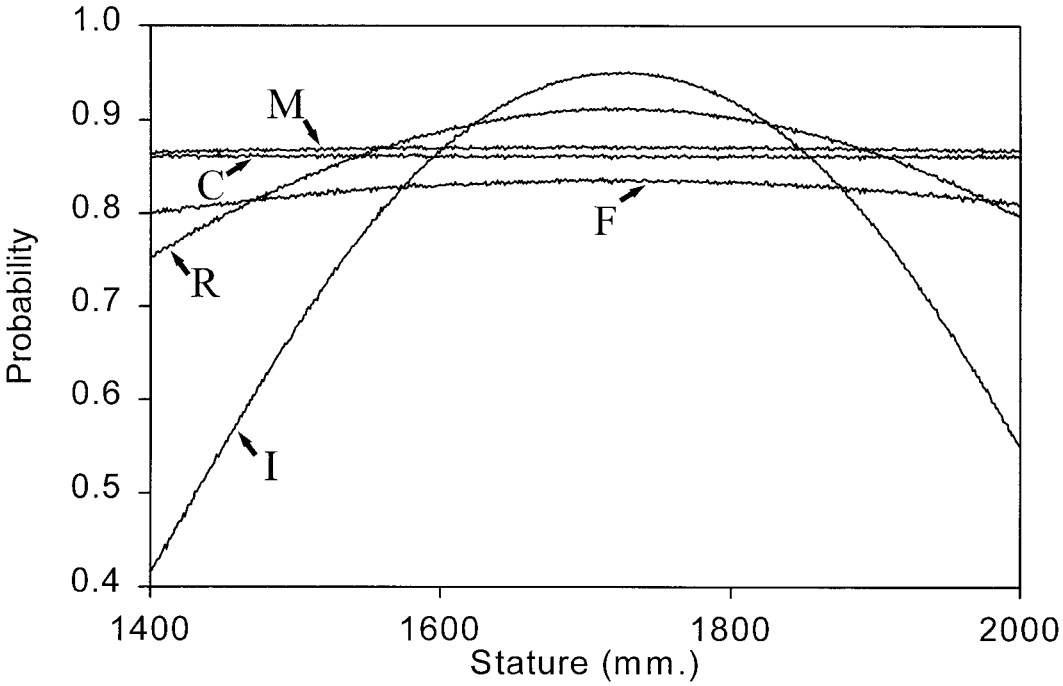

Fig. 9. Monte Carlo comparisons of stochastic dominance for inverse calibration (I), classical calibration (C), reduced major axis (R), major axis (M), and femur/stature ratio (F). The plotted lines across stature are the probabilities that each estimator will be within 72 mm of the true value of stature.

Are we extrapolating when we estimate stature?

Before we turn to a discussion of multivariate-univariate calibration, we should comment on the problem of extrapolation when estimating stature. We have seen that inverse calibration gives quite biased estimates if we are extrapolating beyond the limits of the reference sample stature distribution and if the correlation between a long bone and stature is less than 1.0. This raises the question of whether it is possible to recognize when we estimate stature whether we are interpolating as vs. extrapolating from the reference data. There is no statistical test that can tell us whether we are extrapolating, as any such test would be circular in nature. For example, if we find a femur which appears to be quite long, we might presume from the femur's excessive length that the individual was quite tall. Would this justify using a tall reference sample and inverse calibration? In point of fact, it does not. As the correlation between stature and femur length is imperfect, we cannot know whether the individual had a long femur because he was tall or because she had a large positive deviation from the regression line. Were we to proceed using inverse calibration with a tall reference sample, we would be guilty of using a posteriori information as if it were prior to the analysis. We can contrast this situation with the case in which we know that an individual is from a tall population. In this setting, it is perfectly appropriate to use an informative prior. The simplest way to incorporate this prior is to use inverse calibration (regression of stature on femur length) with a reference sample that is representative of the population from which the individual was drawn. Lwin and Maritz (1982:240) succinctly summarize the problem of interpolation vs. extrapolation, noting that

> In practice our estimation problem is either of interpolation (restricted mainly to the interior of the calibration range) or extrapolation (restricted to the extremities and outside of the calibration range). However, whether

one would be interpolating or extrapolating can be judged only if we have some prior information about x such as a probability distribution of x. In such a bivariate regression the inverse estimator can be useful. Otherwise there seems to be no justification for its use.

We will see that, in contrast to the univariate-univariate problem, in the multivariate-univariate calibration problem there is a diagnostic that can be used to determine whether or not we are extrapolating.

MULTIVARIATE-UNIVARIATE CALIBRATION

Our discussion so far has considered only cases in which there is one *x* variable (stature) and one *y* variable (femoral length). This is essentially a bivariate problem but is often referred to as a univariate-univariate calibration problem because there is one predictor and one variable to be predicted. A more general problem is multivariate-univariate calibration in which two or more variables are used to estimate one unobserved variable for a new case. We consider such calibrations in this section, in which we examine the use of femur and humerus length to estimate stature. For inverse calibration, the multivariate generalization is immediate. As early as 1899, Pearson (1899) used multiple regression of stature on long bones in a reference collection to estimate stature for a target sample. This is the multivariate generalization for inverse calibration. Major axis and reduced major axis can also be extended to the multivariate case (McArdle, 1988), as we discuss below. Additionally, there are other multivariate methods which can be used to estimate stature, such as principal components regression and partial least squares. These and other estimators for multivariate calibration are reviewed in Martens and Næs (1991) and Brown (1993).

The multivariate generalization of classical calibration has been discussed at length in Lieftinick-Koeijers (1988), Mathew and Kasala (1994), and Srivastava (1995). The multivariate classical calibration estimator is

$$\hat{x}_i = (\mathbf{y}_i - \mathbf{a})' C^{-1}\mathbf{b}(\mathbf{b}'\mathbf{C}^{-1}\mathbf{b})^{-1}, \qquad (9)$$

where $\mathbf{y}_i$ is a vector of bone measurements for individual *i* from the target sample, **b** is a vector of regression coefficients for bone measurements on stature in the reference collection, **a** is a vector of y-intercepts for each measurement, and **C** is the variance-covariance matrix among bone measurements after regression on stature in the reference collection. **C**, **b,** and **a** can be obtained by any statistical package capable of performing a multivariate regression of two or more variables on a single "explanatory" variable. Alternatively, a series of univariate regressions of bone measurements on stature can be performed, the residuals can be found for each regression, and then the variance-covariance matrix can be formed from the residuals. This can be done in even the simplest of computer spreadsheet packages.

The multivariate generalization of the femur/stature ratio method follows from equation 8. In this method, the intercepts are all zero, so **a** drops from the equation, and **b** is now a vector of mean bone measurements divided by mean stature in the reference collection (we replace this vector with the symbol **m**, representing a column vector of ratios). **C** remains as the residual variance-covariance matrix among bone measurements after prediction by each bone/stature ratio. This gives the ratio estimator as

$$\hat{x}_i = \mathbf{y}'_i\mathbf{C}^{-1}\mathbf{m}(\mathbf{m}'\mathbf{C}^{-1}\mathbf{m})^{-1}. \qquad (10)$$

The residual variance-covariance matrix (**C**) can be found by explicitly forming the bone/stature ratios in the reference sample, estimating bone measurements from the known statures and ratios in the reference, forming the residuals, and then calculating **C** from the residuals. Alternatively, **C** can be found directly from the sufficient multivariate statistics. We let **V** be the variance-covariance matrix among measurements and partition this matrix into $\mathbf{V}_{11}$, which is a single element containing the variance of stature, $\mathbf{V}_{22}$, which is the variance-covariance matrix among long-bone lengths, $\mathbf{V}_{21}$, which is a column vector of the covariances of stature against bone lengths, and $\mathbf{V}_{12}$, the transpose of $\mathbf{V}_{21}$. The residual variance-covariance matrix from bone/stature

TABLE 4. Regression equations for predicting stature (in millimeters) from maximum humerus and femoral lengths (in millimeters)[1]

Method	Intercept	Humerus	Femur	$\sqrt{\text{IMSE}}$*
Inverse calibration	452.85	1.0602	1.9745	41.10
Classical calibration	69.91	1.3793	2.5687	46.90
RMA	404.65	−4.7484	6.2080	73.01
Ratio	0.0	1.8322	2.3965	48.82
MA	1,978.06	−73.9034	52.0066	742.72

[1] Based on 2,053 humans (Terry, WWII, and Forensic Databank).
*Calculated under the same assumption as in Table 2.

ratio predictions can then be found as

$$\mathbf{C} = \mathbf{V}_{22} + \mathbf{V}_{11}\mathbf{m} \cdot \mathbf{m}' - \mathbf{m} \cdot \mathbf{V}_{12} - (\mathbf{m} \cdot \mathbf{V}_{12})'. \quad (11)$$

McArdle (1988) has presented the multivariate extensions for major axis and reduced major axis regression. For major axis regression, it is necessary to find the eigenvector of the variance-covariance matrix (**V**) associated with the smallest eigenvalue. If we write this as **e**, a column vector where the first element (e_1) refers to stature and the remaining elements ($e_{2..n}$) refer to long bones, then the vector of regression coefficients for stature on long-bone lengths is $-e_{2..n}/e_1$. In reduced major axis regression, it is necessary to find the eigenvector associated with the smallest eigenvalue from the correlation matrix among stature and bone lengths. The regression coefficients are then given as in major axis regression but must be scaled by the ratios of the standard deviation of stature to the standard deviations of each long bone.

While the mechanics of finding the intercept and regression coefficients for each method may appear formidable (Table 4 gives the regressions of stature on humerus and femur), all the methods are simple to understand at an intuitive level. Each method seeks to fit a plane through the trivariate data and then project future bivariate cases (observed femur and humerus length) onto the plane to estimate stature. In inverse calibration (multiple regression of stature on femur and humerus length), the plane is fit so that it minimizes the distances of actual statures to the plane. In classical calibration, the plane is fit so that it simultaneously minimizes the distances of femoral and humeral measurements to the plane. The ratio estimator also fits the plane in this fashion but with the restriction that the plane must pass through the origin (0,0,0). In the multivariate extension of RMA, the plane is fit to simultaneously minimize the distances of femoral, humeral, and statural measurements to the plane. From a geometric standpoint, the trivariate RMA minimizes the volume of tetrahedra, where each tetrahedron has a face in the fitted plane and each of the remaining three faces is parallel to one of the coordinate axes. Finally, in MA the plane is fit to minimize the perpendicular distances of points to the plane (i.e., the lengths of the vectors that project points onto the plane are minimized).

Regardless of the multivariate-univariate method used, the large-sample expected mean squared error is relatively simple to find. All methods require finding the regression coefficients of long bones on stature (which are equal to $\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$, and which we write as β) and the residual variance-covariance matrix among long bones after regression on stature (which are equal to $\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$, and which we write as **C**). If we use β_m as shorthand for the regression coefficients from any particular method (e.g., multivariate MA, RMA, classical or inverse calibration, or the ratio method), then the expected mean square at the true value of stature (x) is

$$MSE_{meth} = [(1 - \beta'\beta_m)(\bar{x} - x)]^2 + \beta'_m\mathbf{C}\beta_m. \quad (12)$$

In this equation, the first term is the squared bias, while the second term is the conditional variance. In classical calibration, the term $1 - \beta'\beta_m$ is equal to zero, which again shows that this method is unbiased. The integrated mean squared error is then

$$\text{IMSE}_{\text{meth}} = (1 - \beta'\beta_m)^2\mathbf{V}_{11} + \beta'_m\mathbf{C}\beta_m. \quad (13)$$

In our humerus and femur example from the 2,053 individuals, the term $1 - \beta'\beta_m$ is positive for all methods except the ratio method (the values are 0.2313, 0.0091, and 0.1839 for inverse calibration, major axis, and reduced major axis regression, respectively). Positive values indicate that the estimator will overestimate stature for indi-

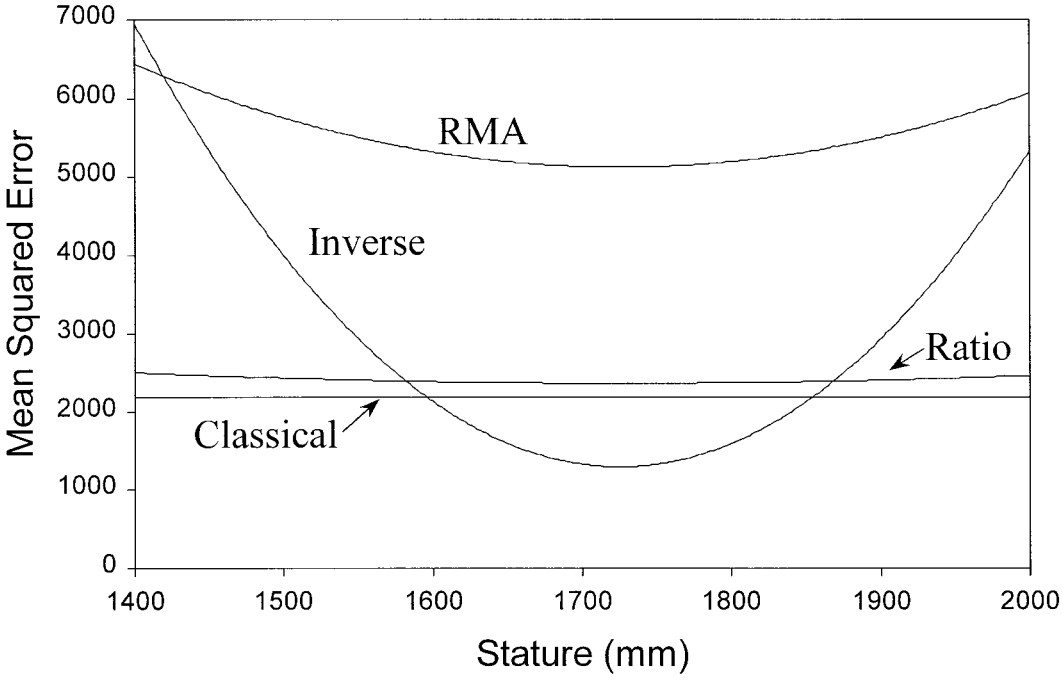

Fig. 10. Theoretical mean squared errors (MSE) for stature estimated from femur and humerus length conditional on true stature. The MSE from major axis is in excess of 543,000 and is not shown here.

viduals who are below the mean stature and underestimate stature for individuals who are above the mean stature. The ratio method has a $1 - \beta'\beta_m$ term that is negative (-.0361), which indicates that the estimator underestimates stature for individuals who are below the mean stature and overestimates stature for individuals who are above the mean. The terms for the major axis regression and ratio methods are near to 0.0, suggesting that the bias is fairly negligible. However, for major axis regression, the conditional standard deviation $\left(\sqrt{\beta'_m \mathbf{C} \beta_m}\right)$ is very high (736 mm vs. 36 mm for inverse calibration), which leads to a very high MSE. The ratio estimator has a higher conditional standard deviation (equal to approximately 49 mm) than does the classical calibration estimator (equal to approximately 47 mm). Combined with the fact that the ratio estimator is slightly biased, this means that the ratio estimator has a higher expected MSE for any value of stature. The relationship among the expected MSEs is shown in Figure 10. Table 5 gives the percent bias as $(1 - \beta'\beta_m) \times 100\%$, and the conditional standard deviations for prediction from femur alone and from femur combined with humerus. These results show the particularly poor performance of MA when both femur and humerus are used to estimate stature.

TABLE 5. Percent bias and conditional standard deviation for stature estimated from femur alone and from femur and humerus

	Prediction from femur		Prediction from femur and humerus	
Method	Percent bias	Conditional standard deviation (mm)	Percent bias	Conditional standard deviation (mm)
Inverse	24.58	36.71	23.13	35.95
Classical	0.00	48.68	0.00	46.78
MA	2.49	47.46	0.91	736.36
RMA	13.15	42.27	18.39	48.40
Ratio	−6.45	51.82	−3.61	48.62

A worked example of stature estimation from the humerus and femur

Here we present a brief worked example of stature estimation from the humerus and femur. For this example, we use the param-

TABLE 6. Root mean square error (RMSE) and bias (both in millimeters) for stature predictions of 159 forensic cases using 1,894 individuals from the Terry and WWII samples for calibration[1]

Method	RMSE	Bias
Inverse calibration	54	−2
Classical calibration	57	−3
Reduced major axis	54	−3
Femur/stature ratio	60	5
Major axis	54	4

[1] Negative bias indicates underestimates, while positive bias indicates overestimates.

eters from Table 4 (based on 2,053 humans) to estimate stature for the shortest Pygmy among our data (taken from Flower, 1888). When this Pygmy's humerus measurement of 243 mm and femoral measurement of 334 mm are entered into the regression equations from Table 4, the estimated stature from inverse calibration is (to the nearest millimeter) 1,370 mm, while from classical calibration the estimated stature is 1,263 mm. RMA and MA give estimates of 1,324 and 1,390, respectively, while the ratio estimate is 1,246. Only the estimates from classical calibration and the ratio estimator are close to the estimated stature (from the Fully method) of 1,211 mm. When the square root of IMSE from Table 4 is used, the 95% confidence intervals around each estimate for inverse calibration, classical calibration, RMA, MA, and the ratio estimator are from 1,289–1,451, from 1,171–1,355, from 1,181–1,467, from 1,150–1,341, and from -66–2,845 mm, respectively. A few comments are in order here concerning these confidence intervals. First, with the exception of classical calibration, the calculation of the confidence intervals assumes that we are not extrapolating. Second, all methods assume that the case for which we are estimating stature follows the same allometry as the reference sample. Later we will discuss tests of these assumptions that are available for multivariate-univariate calibration but not for univariate-univariate calibration.

PERFORMANCE OF STATURE ESTIMATORS

An interpolation example

Table 6 presents the root mean square errors (RMSE) and biases across 159 forensic cases with statures estimated from femoral lengths. Here the stature prediction equations are based on the sample of 1,894 individuals from the Terry and WWII collections, and the statures for the forensic cases come from documented measurements or reported measurements taken during life. As one would expect, the inverse calibration method (regression of stature on femur) gives among the lowest RMSE at 54 mm. Because the means for stature in the forensic sample and the Terry and WWII samples are so similar (see Table 1), we would expect inverse calibration to be the minimum variance unbiased estimator. RMA and MA have, within rounding error, RMSEs equivalent to inverse calibration. In contrast, classical calibration has a RMSE of 57 mm, and the femur/stature ratio has the highest RMSE at 60 mm. As would be expected in a case of interpolation rather than extrapolation, all methods give relatively unbiased estimates.

We have not tabled the simultaneous and pairwise PMCs among the five estimators of stature because this information is voluminous and not very useful. The pairwise PMCs do not suggest that we should have any particularly strong preference among the five estimators, while the simultaneous PMCs indicate that inverse regression is the "best" estimator most frequently (46.5% of the time). At a frequency of 47.2%, this estimator is also the second most probable estimator (after the femur/stature ratio at 52.2%) to give the "worst" estimate. Inverse calibration, and for that matter the femur/stature ratio, is frequently the "best" or the "worst" but rarely intermediate. In the case of inverse calibration, this is true because the method does well for true values near the reference sample mean and poorly for values away from the mean.

For comparison to Table 6, Table 7 presents the RMSEs and biases for the multivariate (femur and humerus) calibration problem. The results are quite similar to the univariate (femur) calibration, with the exception of the MA method, in which the estimator grossly overestimates stature.

An extrapolation example

As it is impossible to obtain a large sample of known stature fossil hominid remains, we

TABLE 7. Root mean square error (RMSE) and bias (both in millimeters) for multivariate (humerus and femur) stature predictions of 159 forensic cases using 1,894 individuals from the Terry and WWII samples for calibration[1]

Method	RMSE	Bias
Inverse calibration	52	−6
Classical calibration	55	−8
Reduced major axis	82	4
Femur/stature ratio	57	−9
Major axis	742	117

[1] Negative bias indicates underestimates, while positive bias indicates overestimates.

TABLE 8. Root mean square error (RMSE) and bias (both in millimeters) for stature predictions (from femora) of 19 Pygmies and the shortest 180 individuals from the Terry/WWII/forensic sample using 1,873 individuals from the Terry, WWII, and forensic samples for calibration[1]

Method	RMSE	Bias
Inverse calibration	81	68
Classical calibration	50	2
Reduced major axis	59	38
Femur/stature ratio	54	−11
Major axis	50	10

[1] Negative bias indicates underestimates, while positive bias indicates overestimates.

have instead opted to use the 180 shortest individuals (1,390–1,600 mm) from the combined Terry, WWII, and forensic data as a test set for calibrations from the taller individuals from this combined data set. We also include the 19 African Pygmies in this test set to bring the total to 199 individuals. The remaining 1,873 individuals from the Terry, WWII, and forensic data are used as the calibration sample to estimate statures for the 199 test cases.

Table 8 shows the RMSEs and biases for the five estimators. As would be expected in a case of extrapolation such as this, inverse calibration is quite biased, leading to an average overestimation for the short sample of 6.8 cm. Reduced major axis, as it is intermediate between the biased inverse calibration and the (theoretically) unbiased classical calibration, also gives biased estimates. Reduced major axis overestimates on average by 3.8 cm. Major axis overestimates trivially by 1.0 cm on average, while the femur/stature ratio is the only estimator which leads to underestimates (1.1 cm on average). As one would predict in such a case of extrapolation, classical calibration is relatively unbiased, overestimating on average by 2 mm. The RMSEs show that inverse calibration has an unacceptably large level of error (8.1 cm) compared to the other estimators, again because of its bias. Classical calibration and major axis regression are tied for the lowest RMSE at 5.0 cm, while the femur/stature ratio has a slightly higher RMSE of 5.4 cm and reduced major axis has again a slightly higher RMSE at 5.9 cm.

TABLE 9. Root mean square error (RMSE) and bias (both in millimeters) for multivariate (femur and humerus) stature predictions of 19 Pygmies and the shortest 180 individuals from the Terry/WWII/forensic sample using 1,873 individuals from the Terry, WWII, and forensic samples for calibration[1]

Method	RMSE	Bias
Inverse calibration	74	62
Classical calibration	46	0
Reduced major axis	84	38
Femur/stature ratio	48	−8
Major axis	566	80

[1] Negative bias indicates underestimates, while positive bias indicates overestimates.

The pairwise PMCs (again, not presented here) indicated that inverse calibration is the Pitman farthest estimator in all comparisons, while reduced major axis regression is the Pitman farthest estimator in all comparisons except against inverse calibration. The pairwise PMCs also show that classical calibration is the closest estimator in all comparisons. The simultaneous PMCs lead to some disparate conclusions from those provided by the pairwise PMCs. The simultaneous "best" estimator is the femur/stature ratio, which was the pairwise "worst" estimator for all comparisons except against inverse regression and reduced major axis regression. The "pairwise-worst simultaneous-best paradox" is well-known and described for Pitman's measure of closeness (Keating et al., 1993), as are a number of other paradoxes.

For comparison to Table 8, Table 9 presents the RMSEs and biases for the multivariate (femur and humerus) calibration problem. Again, the results are similar to the univariate (femur) calibration, with the exception of the MA method, in which the estimator grossly overestimates stature. Hens (1998) provides a much more detailed and extensive discussion of both univariate (femur)

TABLE 10. Tests for extrapolation and allometric differences from the humerus and femur[1]

Specimen	Probable stature	Humerus	Femur	Inverse estimate	Classical estimate	R_x	R
Pygmy	1,211?	243	334	1,370	1,263	22.60[2]	0.03
A.L. 288-1	1,050?	239	281	1,261	1,121	38.56[2]	8.94[2]
KNM-WT 15000	1,600?	319	432	1,644	1,620	1.18	1.04
La Ferrasie 1	1,690?	340	460	1,722	1,720	0.00	1.66
La Ferrasie 2	1,550?	286	411	1,568	1,520	4.45[2]	0.89

[1] Bone measurements and probable statures are from the following sources: Pygmy: Flower (1888), Jungers (1982); A.L. 288-1: Jungers (1982), Geissman (1986); KNM-WT 15000: Ruff and Walker (1993); La Ferrasie 1 and 2: Wolpoff (1999).
[2] Significant at the 0.05 level on a chi-square test with one degree of freedom.

and multivariate (femur and humerus) estimation of stature for hominids.

Extrapolation, consistency, and confidence in multivariate-univariate calibration

We close our presentation of basic results with a consideration of extrapolation in the multivariate-univariate calibration setting. Unlike the univariate-univariate case, there is now the possibility for testing whether there is extrapolation and whether the long bones give consistent information about stature. Our summary of methods here follows Brown and Sundberg (1987) and Brown (1993).

If we use multiple regression of stature on long bones (i.e., multivariate-univariate inverse calibration) to estimate the stature for a new case, then we can follow this by predicting the long-bone lengths that should occur in someone with that estimated stature. Brown (1993) gives the prediction covariance matrix among the observed and predicted long-bone lengths as

$$E(y_0 - \check{y})(y_0 - \check{y})' = C(C + \beta \mathbf{V}_{11} \beta')^{-1} C, \quad (14)$$

where y_o is the vector of observed long-bone lengths and $\check{y}$ is the vector of long-bone lengths predicted by regressing stature on long bones to estimate stature and then regressing long bones on stature to estimate the (observed) long-bone lengths. In equation 14, β is again the vector of regression coefficients of long bones on stature. Brown (1993) defines a matrix Θ as the inverse of the prediction covariance matrix in equation 14. This matrix forms the basis for two useful quadratic forms. One of the forms, which Brown (1993) calls R_X, is

$$R_X = (\hat{y} - \check{y})' \Theta (\hat{y} - \check{y}), \quad (15)$$

where $\hat{y}$ is a vector of predicted long-bone lengths estimated by using classical calibration to estimate stature and the regressions of long-bone's lengths on stature to obtain the long bone lengths. R_X is asymptotically distributed as a chi-square with one degree of freedom. As Brown (1993) notes, R_X increases as the actual stature departs from the mean stature of the reference sample (because the classical and inverse estimates of stature diverge). The second quadratic Brown (1993) writes as "R," and Brown and Sundberg (1987:49) refer it to as an "inconsistency diagnostic." This statistic is

$$R = (y_0 - \hat{y})' \Theta (y_0 - \hat{y}). \quad (16)$$

R is asymptotically distributed as a chi-square with $\ell - 1$ degrees of freedom, where ℓ is the number of long bones. Within the context of stature estimation, R is a measure of departure from allometric scaling. If the case for which stature is being estimated differs allometrically from the reference sample, then R will be large. In essence, R_X tests for size difference, while R tests for shape difference between the reference sample and the case for which we are estimating stature.

Table 10 presents a few examples of R_X and R calculated from the shortest Pygmy in our sample as well as from A.L. 288–1, KNM-WT 15000 (Ruff and Walker, 1993), and La Ferrassie 1 and 2 (Wolpoff, 1999). To construct this table, we used the calibration set of 2,053 non-Pygmy humans. For comparison, Table 10 also presents the statures that would be estimated by multivariate

inverse and classical calibration from the femur and humerus. The R_X statistic indicates that the Pygmy, A.L. 288–1, and La Ferrasie 2 all represent significant extrapolations from the modern non-Pygmy human sample (with its mean stature of 1,725 mm). The R statistic shows that only A.L. 288–1 differs from the modern human allometric pattern for the humerus and femur. Taken as a whole, these results suggest that we should use classical calibration to estimate the Pygmy's and La Ferrassie 2's statures from the humerus and femur, while we can use inverse calibration to estimate statures for KNM-WT 15000 and La Ferrasie 1. For A.L. 288–1, neither classical nor inverse multivariate calibration can be used. Instead, we would have to choose among the classical calibration estimates from either the humerus or the femur. Because of the femur's biomechanical relevance (as well as its direct contribution to stature), the femoral estimate is the obvious choice.

We have been noticeably mute on the subject of confidence intervals in multivariate-univariate calibration problems. This is a difficult subject and one that we do not feel we can cover in any great depth here. In Table 4 we have listed the square root of the integrated mean squared error (IMSE) for each method, and in theory we could use plus and minus 1.96 times these values around estimated statures to give 95% confidence intervals. The IMSEs were calculated using equation 13 and agree with the multivariate forms for inverse and classical multivariate calibration given in Brown (1979). While we know that the calculation of these values is correct, the assumptions underlying the calculations may not be. To use the IMSE for inverse calibration, we have already assumed that extrapolation is very unlikely (because the integration is across the reference sample stature distribution), while for all methods we have ignored the problems that arise when allometry differs (i.e., R is not equal to zero). Brown (1993) and Brown and Sundberg (1987) note that in multivariate classical calibration the widths of confidence intervals actually decrease as R increases. This is an altogether undesirable property of the classical estimator, because increasing R should be reflected in increasing uncertainty about stature. However, Brown and Sundberg (1987) show that the classical estimator is not the maximum likelihood estimator in multivariate calibration. They show using a profile likelihood argument that the maximum likelihood estimator (i.e., the stature at which the observed bone lengths are most likely to have occurred given the reference sample data) does provide a confidence interval which increases with increasing R. Du Plessis and van der Merwe (1995) give a multivariate Monte Carlo Bayesian estimator that could be used to find the posterior distribution for stature conditional on the reference sample information and observed long-bone lengths from an individual. Their method uses a diffuse prior (i.e., relatively uninformative) and as a consequence should differ little from a maximum likelihood method. Huet et al. (1996) describe a bootstrap for calibration intervals that is in the spirit of the classical estimator and would be useful for small samples. Unfortunately, their method is described only for univariate calibration.

DISCUSSION

Our examination of MSE, PMC, and stochastic domination for five stature estimators provides a clear message regarding preferences for stature-estimation methods. When it may be argued that an individual case derives from the same stature distribution as the reference sample, inverse calibration is clearly the preferred method. In the vicinity of the mean stature, regression of stature on one or more long-bone lengths will provide the lowest MSE and highest PMC. The integrated mean square error is the lowest for inverse calibration, provided the integration is truly across the reference sample stature distribution. Generally speaking, in paleoanthropology or bioarchaeology, where it may not be possible to know that individual specimens or samples come from the same stature distribution as the reference sample, the regression of stature on one or more long-bone lengths may produce very biased estimates. Although RMA has been suggested as an alternative approach in these contexts (Aiello, 1992; Sjøvold, 1990), this method is essentially a compromise between inverse calibration (re-

gression of stature on long bone) and classical calibration (regression of long bone on stature followed by solving for stature). Because it is a compromise, RMA extends the useful range around the mean stature (i.e., allows further extrapolation) relative to inverse calibration but not to the extent of classical calibration. As we may well find ourselves in situations where extreme extrapolation is necessary, even RMA may be inappropriate. In such settings, our choice must be among classical calibration, the femur/stature ratio, and MA.

In cases of extrapolation, the choice between classical calibration and the femur/stature ratio is quite clear. The asymptotic PMC and MSE results both indicated that the femur/stature ratio was an inadmissible stature estimator when compared to classical calibration. The reason for this is that the femur/stature ratio is a special case of classical calibration. Specifically, when the femur scales isometrically against stature, the femur/stature ratio becomes the classical calibration estimator. In order to demonstrate this, we need a definition for isometry. Jungers et al. (1995:139) give a general definition stating that isometry "implies that shape is preserved among organisms of different sizes. In raw data space, points lying on the same positively directed ray emanating from the origin exhibit this quality." As we have taken stature to be size here, the "positively directed ray" is the least-squares regression line of femur length on stature. In this context, the definition of isometry is that the ratio of the coefficient of variation of stature (CV_x) to the coefficient of variation of femur length (CV_y) is equal to the Pearson correlation (r) between femur length and stature ($CV_x / CV_y = r$). This definition is (necessarily) different from some other definitions of isometry which differ in how size itself is defined (e.g., see results from Cole, 1995). Solving for the regression of femur length on stature under the current definition of isometry gives a regression equation with an intercept of zero and a slope equal to the average femur length divided by average stature. In other words, the regression line for femur length on stature coincides with the femur/stature ratio line, so the estimates of stature from these two identical lines must be identical. A more formal proof that the femur/stature ratio is a special case of classical calibration is given in Hens et al. (1998). If the femur does not scale isometrically with stature, then the femur/stature ratio will produce a biased estimator for stature. This has previously been pointed out by Meadows and Jantz (1995:765–766), who wrote

> Our results showing that the femur is positively allometric indicate that the femur-stature ratio varies with height. Using a constant ratio, we expect that taller individuals, with relatively longer femora would be overestimated.

It is also possible to show that under isometry classical calibration (and therefore the femur/stature ratio as well) operates as a shape variable in the sense used by Mosimann (1970). Welsh et al. (1988) show that the expected value for the ratio of two variables from a bivariate normal distribution is

$$E\left[\frac{y}{x}\right] = \frac{\bar{y}}{\bar{x}} - \frac{\operatorname{cov}\left(\frac{y}{x}, x\right)}{\bar{x}}, \qquad (17)$$

where the expected value for the ratio (on the left-hand side of the equation) is an average Mosimann shape variable, and the ratio of means on the right-hand side could be interpreted as the femur/stature ratio. The covariance of a ratio with its denominator is zero when $CV_x / CV_y = r$, as in our definition of isometry (this result can be found from Chayes's (1949) equation 8). Consequently, under isometry the average ratio will be equal to the ratio of the averages, so the femur/stature ratio is an average Mosimann shape variable.

As the femur/stature ratio cannot be recommended over classical calibration either in cases of extrapolation or interpolation, the only remaining comparison is between classical calibration and major axis regression. None of the criteria we have examined (MSE, PMC, and stochastic domination) lead to a clear choice between classical calibration and MA in the univariate-univariate calibration problem. We are instead struck by the similarity between the two estimators. In the univariate-multivariate calibration problem (i.e., stature estimated from multiple bones) MA performs remarkably

poorly relative to all other estimators. This occurs because of the assumption of equal error variance for humerus, femur, and stature in the multivariate MA. Based on the very high MSE for multivariate MA, we cannot recommend the use of MA.

CONCLUSIONS

We recommend using the regression of stature on bone length(s) when there is some a priori reason for presuming that a case comes from the same stature distribution as represented within the reference (calibration) sample. Such a presumption is generally warranted only in the forensic setting and in paleoanthropology when there is some theoretical reason for a strong prior. In either case, the reference sample will need to be representative of the individual case (i.e., the individual case should be an interpolation, not an extrapolation). Even in the forensic setting it may not be possible to make this assumption if the reference sample represents a population or populations which differ markedly from potential forensic cases or if the sampling methods differ markedly between the reference sample and the forensic cases. If we are unwilling to assume that a case or specimen comes from the same stature distribution as the reference sample, then classical calibration (the regression of bone length(s) on stature, followed by solving for stature) should be used. Classical calibration comes with the cost that confidence intervals around estimated statures will be broader. We see this as a small price to pay and preferable to the embarrassment of providing tighter standard errors around biased estimates from inverse calibration.

It is in some sense ironic that we find ourselves primarily recommending classical calibration as the method of choice for estimating statures in some contexts within physical anthropology. Common practice in virtually all other fields has been to use classical calibration, and it was not until Krutchkoff's (1967) recommendation of inverse calibration that a vituperative argument erupted (reviewed in Osborne, 1991). In contrast, since as early as the late nineteenth century (Pearson, 1899), inverse calibration has been used to estimate human stature. We now find ourselves in the unenviable position of recommending classical calibration, even at the risk of another lengthy exchange erupting, such as occurred in the late 1960s. The more remarkable point is that we are recommending a maximum likelihood approach, at least in some contexts, over a very entrenched Bayesian approach. While a debate wages in the statistical literature over the merits of teaching Bayesian statistics in introductory classes (Albert, 1997; Berry, 1997; Moore, 1997), we find that physical anthropologists have long been using a Bayesian approach with little awareness of its inferential basis. Perhaps the oft perceived ills of Bayesian inference can be tempered with the realization that many of us have (unwittingly) made such inferences in the past.

LITERATURE CITED

Afifi AA, and Clark V (1996) Computer-Aided Multivariate Analysis, 3rd ed. New York: Chapman and Hall.
Aiello LC (1992) Allometry and the analysis of size and shape in human evolution. J. Hum. Evol. *22:*127–147.
Albert J (1997) Teaching Bayes' rule: A data-oriented approach. Am. Statist. *51:*247–253.
Aykroyd RG, Lucy D, Pollard AM, and Solheim T (1997) Regression analysis in adult age estimation. Am. J. Phys. Anthropol. *104:*259–265.
Berkson J (1969) Estimation of a linear function for a calibration line: Consideration of a recent proposal. Technometrics *11:*649–660.
Berry DA (1997) Teaching elementary Bayesian statistics with real applications in science. Am. Statist. *51:*241–246.
Bissell AF (1992) Lines through the origin—is NO INT the answer? J. Appl. Statist. *19:*193–210.
Blyth CR (1993) Comment on Robert, Hwang, and Strawderman (1993). J. Am. Statist. Assoc. *88:*72–74.
Boldsen J (1984) A statistical evaluation of the basis for predicting stature from lengths of long bones in Europeans. Am. J. Phys. Anthropol. *65:*305–311.
Bolfarine H, and Lima CROP (1997) Linear calibration in functional regression models. Commun. Statist. Theory Meth. *26:*2307–2328.
Brown GH (1979) An optimization criterion for linear inverse estimation. Technometrics *21:*575–579.
Brown PJ (1982) Multivariate calibration. J. R. Statist. Soc. B *44:*287–321.
Brown PJ (1993) Measurement, Regression, and Calibration. Oxford: Oxford University Press.
Brown PJ, and Sundberg R (1987) Confidence and conflict in multivariate calibration. J. R. Statist. Soc. B *49:*46–57.
Brownlee KA (1965) Statistical Theory and Methodology in Science and Engineering. New York: John Wiley and Sons.
Carroll RJ, Spiegelman CH, and Sacks J (1988) A quick and easy multiple-use calibration-curve procedure. Technometrics *30:*137–141.
Casella G, and Wells MT (1993) Comment on Robert, Hwang, and Strawderman (1993). J. Am. Statist. Assoc. *88:*70.

Chayes F (1949) On ratio correlation in petrography. J. Geol. *57:*239–254.
Cole TM III (1995) Comparative Craniometry of the *Atelinae (Platyrrhini, Primates)*: Function, Development, and Evolution. Ph.D. dissertation, SUNY Stony Brook.
Du Plessis JL, and van der Merwe (1995) A Bayesian approach to multivariate and conditional calibration. Computational Statist. Data Anal. *19:*539–552.
Eisenhart C (1939) The interpretation of certain regression methods and their use in biological and industrial research. Ann. Math. Statist. *10:*162–186.
Feldesman MR, and Fountain RL (1996) "Race" specificity and the femur/stature ratio. Am. J. Phys. Anthropol. *100:*207–224.
Feldesman MR, and Lundy JK (1988) Stature estimates for some African Plio-Pleistocene fossil hominids. J. Hum. Evol. *17:*583–596.
Feldesman MR, Kleckner JG, and Lundy JK (1990) Femur/Stature ratio and estimates of stature in Mid- and Late-Pleistocene fossil hominids. Am. J. Phys. Anthropol. *83:*359–372.
Fisch RD, and Strehlau GA (1993) A simplified approach to calibration confidence sets. Am. Statist. *47:*168–171.
Flower WH (1888) Description of two skeletons of Akkas, a pygmy race from Central Africa. J. Roy. Anthropol. Inst. *18:*3–18.
Fully G (1956) Une nouvelle methode de determination de la taille. Ann. Med. Leg. Criminol. *35:*266–273.
Geissmann T (1986) Estimation of Australopithecine stature from long bones: A.L. 288–1 as a test case. Folia Primatol. *47:*119–127.
Gelman A, Carlin JB, Stern HS, and Rubin DB (1995) Bayesian Data Analysis. London: Chapman and Hall.
Giles E, and Hutchinson DL (1991) Stature- and age-related bias in self-reported stature. J. Forensic Sci. *36:*765–780.
Giles E, and Klepinger LL (1988) Confidence intervals for estimates based on linear regression in forensic anthropology. J. Forensic Sci. *33:*1218–1222.
Gilks WR, and Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Appl. Statist. *41:*337–348.
Graybill FA (1961) An Introduction to Linear Statistical Models. New York: McGraw-Hill.
Green PJ (1981) Peeling bivariate data. In V Barnett (ed.): Interpreting Multivariate Data. New York: John Wiley and Sons, pp. 3–19.
Halperin M (1970) On inverse estimation in linear regression. Technometrics *12:*727–736.
Hens SM (1998) Stature Estimation in Fossil Hominids. Ph.D. dissertation, University of Tennessee, Knoxville.
Hens SM, Konigsberg LW, and Jungers WL (1998) Estimation of African ape body length from femur length. J. Hum. Evol. *34:*401–411.
Hoadley B (1970) A Bayesian look at inverse linear regression. J. Am. Statist. Assoc. *65:*356–369.
Holcomb SMC, and Konigsberg LW (1995) Statistical study of sexual dimorphism in the human fetal sciatic notch. Am. J. Phys. Anthropol. *97:*113–125.
Holliday TW, and Ruff CB (1997) Ecogeographical patterning and stature prediction in fossil hominids: Comment on M.R. Feldesman and R.L. Fountain. Am. J. Phys. Anthropol. *103:*137–140.
Huet S, Bouvier A, Gruet M-A, and Jolivet E (1996) Statistical Tools for Nonlinear Regression. New York: Springer-Verlag.
Hunter WG, and Lamboy WF (1981) A Bayesian analysis of the linear calibration problem. Technometrics *23:*323–350.
Hwang JT (1985) Universal domination and stochastic domination: Estimation simultaneously under a broad class of loss functions. Ann. Statist. *13:*295–314.
Jantz LM (1996) Secular Change and Allometry in the Long Limb Bones of Americans from the Mid 1700's Through the 1900's. Ph.D. dissertation, University of Tennessee, Knoxville.
Jolicoeur P (1968) Interval estimation of the slope of the major axis bivariate normal distribution in the case of a small sample. Biometrics *24:*679–682.
Jolicoeur P (1990) Bivariate allometry: Interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. J. Theor. Biol. *144:*275–285.
Jungers WL (1982) Lucy's limbs: Skeletal allometry and locomotion in *Australopithecus afarensis*. Nature *297:*676–678.
Jungers WL (1988) Lucy's length: Stature reconstruction in *Australopithecus afarensis* (A.L. 288–1) with implications for other small-bodied hominids. Am. J. Phys. Anthropol. *76:*227–231.
Jungers WL, Falsetti AB, and Wall CE (1995) Shape, relative size, and size-adjustments in morphometrics. Yearbook Phys. Anthropol. *38:*137–161.
Keating JP, and Mason RL (1991) Closeness comparison of classical and inverse regression estimators. Comput. Statist. *12:*4–11.
Keating JP, Mason RL, and Sen PK (1993) Pitman's Measure of Closeness: A Comparison of Statistical Estimators. Philadelphia: Society for Industrial and Applied Mathematics.
Konigsberg LW (1991) An historical note on the t-test for differences in sexual dimorphism between populations. Am. J. Phys. Anthropol. *84:*93–96.
Konigsberg LW, and Frankenberg SR (1994) Paleodemography: "Not quite dead." Evol. Anthropol. *3:*92–105.
Konigsberg LW, Frankenberg SR, and Walker RB (1997) Regress what on what? Paleodemographic age estimation as a calibration problem. In RR Paine (ed.): Integrating Archaeological Demography: Multidisciplinary Approaches to Prehistoric Populations. Carbondale, IL: Center for Archaeological Investigation, Occasional Paper 24, pp. 64–88.
Krutchkoff RG (1967) Classical and inverse regression methods of calibration. Technometrics *9:*425–439.
Lee JJ (1991) A note on the conditional approach to interval estimation in the calibration problem. Biometrics *47:*1573–1580.
Lieftinik-Koeijers CAJ (1988) Multivariate calibration: A generalization of the classical estimator. J. Multivariate Analysis *25:*31–44.
Little RJA, and Rubin DB (1987) Statistical Analysis with Missing Data. New York: John Wiley & Sons.
Litton CD, and Buck CE (1995) The Bayesian approach to the interpretation of archaeological data. Archaeometry *37:*1–24.
Lucy D, Aykroyd RG, Pollard RM, and Solheim T (1996) A Bayesian approach to adult human age estimation from dental observations by Johanson's age changes. J. Forensic Sci. *41:*189–194.
Lwin T, and Maritz JS (1982) An analysis of the linear-calibration controversy from the perspective of compound estimation. Technometrics *24:*235–242.
Martens H, and Næs T (1991) Multivariate Calibration. New York: John Wiley and Sons.
Martin RD, and Barbour AD (1989) Aspects of line-fitting in bivariate allometric analyses. Folia Primatol. *53:*65–81.
Martinelle S (1970) On the choice of regression in linear calibration. Comments on a paper by R.G. Krutchkoff. Technometrics *12:*157–161.

Mathew T, and Kasala S (1994) An exact confidence region in multivariate calibration. Ann. Statist. *22:*94–105.
McArdle BH (1988) The structural relationship: Regression in biology. Can. J. Zool. *66:*2329–2339.
Meadows L, and Jantz RL (1995) Allometric secular change in the long bones from the 1800s to the present. J. Forensic Sci. *40:*762–767.
Moore DS (1997) Bayes for beginners? Some reasons to hesitate. Am. Statist. *51:*254–261.
Moore-Jansen PM, Ousley SD, and Jantz RL (1994) Data Collection Procedures for Forensic Skeletal Material. Knoxville: Forensic Anthropology Center, University of Tennessee.
Mosimann JE (1970) Size allometry: Size and shape variables with characterizations of the lognormal and generalized gamma distributions. J. Am. Statist. Assoc. *65:*930–945.
Neter J, Wasserman W, and Kutner M (1985) Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs, 2nd ed. Homewood, IL: Irwin.
Osborne C (1991) Statistical calibration: A review. Int. Statist. Rev. *59:*309–336.
Pearson K (1899) Mathematical contributions to the theory of evolution. V. On the reconstruction of the stature of prehistoric races. Philos. Trans. R. Soc. Lond. *192:*169–244.
Rayner JMV (1985) Linear relations in biomechanics: The statistics of scaling functions. J. Zool. Lond. *206:*415–439.
Ricker WE (1973) Linear regression in fishery research. J. Fish. Res. Board Can. *30:*409–434.
Robert CP, Hwang JTG, and Strawderman WE (1993) Is Pitman closeness a reasonable criterion? J. Am. Statist. Assoc. *88:*57–63.
Rogers NL (1996) A Study of Histological Aging of the Human Clavicle. M.A. thesis, University of Tennessee, Knoxville.
Rosenblat JR, and Spiegelman CH (1981) Discussion of Hunter and Lamboy's 1981 paper. Technometrics *23:* 329–333.
Ruff CB (1994) Morphological adaptation to climate in modern and fossil hominoids. Yearbook Phys. Anthropol. *37:*65–107.
Ruff CB, and Walker A (1993) Body size and body shape. In A Walker and RF Leakey (eds.): The Nariokotome *Homo erectus* Skeleton. Cambridge: Harvard University Press, pp. 234–263.
Schéffe H (1973) A statistical theory of calibration. Ann. Statist. *1:*1–37.
Seim E, and Sæther B-E (1983) On rethinking allometry: Which regression model to use? J. Theor. Biol. *104:*161–168.
Shukla GK (1972) On the problem of calibration. Technometrics *14:*547–553.
Søvold T (1990) Estimation of stature from long bones utilizing the line of organic correlation. Hum. Evol. *5:*431–447.
Smith RJ (1994) Regression models for prediction equations. J. Hum. Evol. *26:*239–244.
Sokal RR, and Rohlf FJ (1981) Biometry, 2nd ed. San Francisco: W.H. Freeman.
Sprent P (1969) Models in Regression and Related Topics. London: Methuen.
Srivastava MS (1995) Comparison of the inverse and classical estimators in multi-univariate linear calibration. Commun. Statist. Theory Meth. *24:*2753–2767.
Stuart A, and Ord JK (1991) Kendall's Advanced Theory of Statistics, Vol. 2, 5th ed. New York: Oxford University.
Trotter M, and Gleser GC (1952) Estimation of stature from long bones of American whites and Negroes. Am. J. Phys. Anthropol. *10:*463–514.
Welsh AH, Peterson AT, and Altmann SA (1988) The fallacy of averages. Amer. Natur. *132:*277–288.
Willey P, and Falsetti T (1991) Inaccuracy of height information on driver's licenses. J. Forensic Sci. *36:* 813–819.
Williams EJ (1969) A note on regression methods in calibration. Technometrics *11:*189–192.
Wolpoff MH (1999) Paleoanthropology, 2nd ed. New York: McGraw-Hill.